MECHANICS OF JUTSUS

THE PLAYER, THE GAME, AND THE JUTSU
MEASURING MECHANICAL EXPERIENCES OF GAMEPLAY.

BY
SASHA M. SORAINE, M.ASc., (Software Engineering)

A THESIS
SUBMITTED TO THE DEPARTMENT OF COMPUTING AND SOFTWARE
AND THE SCHOOL OF GRADUATE STUDIES
OF MCMASTER UNIVERSITY
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTORATE OF SOFTWARE ENGINEERING

Doctorate of Software Engineering (2024)
(Computing and Software)

McMaster University
Hamilton, Ontario, Canada

TITLE:	The Player, the Game, and the Jutsu Measuring mechanical experiences of gameplay.
AUTHOR:	Sasha M. Soraine M.ASc., (Software Engineering) McMaster University, Hamilton, Canada
SUPERVISOR:	Dr. Jacques Carette
NUMBER OF PAGES:	xxi, 242

Lay Abstract

Games entertain us both as players and spectators; but the experiences of playing and watching are vastly different. Like driving a car, *playing* a game requires skills, engagement with the controls, and interaction with the environment. *Watching* is like being a passenger, you can enjoy the scenery, play with the radio, and even feel road rage. But as the passenger doesn't know what it's like to press the pedals, the spectator doesn't have the same experience as the player. We want to understand both experiences; what makes them different, and how does this difference affect their overall game experience? We develop the *Experiential Tetrad* (ExperT) — an experience-type framework that highlights the influence of perspective differences. We then look closer at mechanical experiences from a player perspective, and model its effects with *jutsus* — a visual measure of how a player's skills affect their gameplay experience. We focus our work on developing models and measurements of player abilities and gameplay challenges. This results in three complete jutsus for button mashing challenges. Our goal is that designers can use these as tools to explore the effects of design decisions on experiences, both conceptually with ExperT and practically with jutsus.

Abstract

Research on game-related experiences centres *player experiences* (PX). However, our understanding of PX is limited. At a definitional level, there is no consensus on the dimensions of PX, its proposed constructs overlap in scope, and their specific relationships to game design decisions can be unclear. At a functional level, PX focuses on a narrow set of positive, optimal experiences, even though negative experiences and spectator experiences are also fundamentally important to understanding games. We synthesize various works into the Experiential Tetrad (ExperT) — a perspective-dependent theoretical framework for all game-related experiences. ExperT proposes that holistic experiences are the combination of four interrelated experiential types (mechanical, emotional, aesthetic, and socio-cultural) which are clearly scoped to specific game elements, and timescale oriented. ExperT highlights how the user perspective (e.g. player, spectator) influences the overall experience by changing its characterising experiential type. PX is grounded during play, and so it is heavily characterised by mechanical experiences. Spectator experiences (SX) are grounded out of play, and so are heavily characterised by socio-cultural experiences.

Our work explores mechanical experiences (MX) through a player lens, with the goal of modeling and measuring it. We define MX as the relationship between the player's abilities and the game's challenges. We scope our work to focus on two qualities of MX: *mechanical achievability* (can the player complete the challenge?) and *mechanical difficulty* (how design decisions affect the mechanical achievability). This exploration results in *jutsus* — a knowledge capture artifact that visualizes the MX of a challenge for a particular player. Jutsus connect the challenge's design to the MX, and so could be a useful tool for designers and researchers in understanding a part of PX.

To arrive at jutsus, we construct:

- a Player Model, that describes players by their ability proficiencies (represented as a *player profile*);
- a Challenge Model, that describes gameplay challenges by their ability requirements (represented as a *competency profile*); and,
- a process of comparing player profiles and competency profiles to quantify their MX.

This thesis adds to the fundamental science of games user research (GUR) and human-computer interaction (HCI). It establishes a new theory and opens up new avenues of future work in experience modeling.

To Kalel and Rhael
If you're serious enough about games, you can turn it into a degree.

Acknowledgements

Thank you to my supervisor, Dr. Jacques Carette, for your unfailing support while completing this thesis. You were always willing to indulge my interests, but were sensible enough to reign them in. Without your good judgment, this work would be much longer than it already is.

To my supervisory committee, Dr. Lennart Nacke and Dr. Richard Paige, I am incredibly grateful for your confidence in my ability to finish this marathon of a thesis. Your guidance about scoping was invaluable.

To my external examiner, Dr. Nick Graham. Thank you for reading this monster of a thesis. Your feedback was incredibly useful at presenting my work and considering ways to move forward.

Thank you to my masters students, Xin and Vansh, it was a pleasure supervising you. In guiding your projects, I was able to see my own work in new ways. To Vansh particularly, I am forever grateful in the ways your mini-game project enabled my validation experiments.

To my fellow G-ScalE members, and honorary G-ScalE members: thank you for the years of being my test bed for ideas, good and bad. Particularly to Geneva — nothing brings people closer than the trauma-bond of a doctoral program. Thank you for being my collaborator, sounding board, practice participant, and proofreader.

To my family, thank you for introducing me to games. From sibling board game nights to summers playing video games in the basement, you encouraged and enabled my passion for this medium. Thank you for listening patiently to me explain rules, monologue about what I am playing, or practice research talks. I hope this work makes you proud.

To my loving partner, Aaron, thank you for believing in me. Your unending support is what sustained me through these years. You've seen this thesis from initial conception in my masters through to this moment. I'm so excited to explore the world and research games with you by my side.

We did it team!
Sasha Soraine

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Player experience (PX) is a multi-dimensional [4, 200, 204, 403, 415], context-dependent [319] phenomena resulting from the complex relationship between player and game. PX encompasses many distinct types of experiences based on how the player and game interact mechanically (through gameplay), emotionally (through story), aesthetically (through visuals and sound), and socio-culturally (through the community of players). Modeling these interactions (i.e. mechanical, aesthetic, emotional, and socio-cultural) as experiential types (ExpTypes), may show how these individual experiences construct a holistic PX. We focus on understanding mechanical experience (MX) as a starting point to modeling ExpTypes and PX.

MX comes from the player interacting with mechanically moderated elements of the game, particularly gameplay. Using Adams [7] model of the player-game relationship (Fig. 1.1), gameplay is the combination of *challenges* and in-game *actions*. Actions are an interpretation of inputs made through a controller; ergo, actions are the player's abilities mediated by the controller. **Thus, MX is the relationship between the player's abilities and the gameplay challenges**.

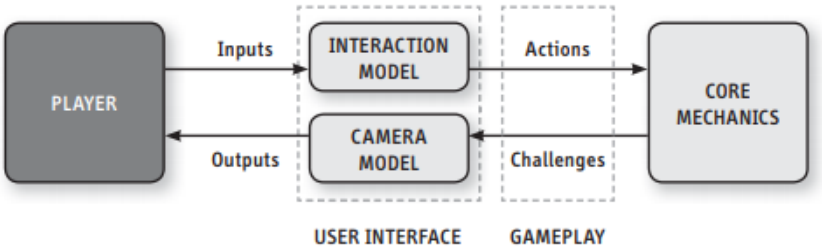


Figure 1.1: [7] model of the game player relationship.

Previous work on a holistic theory of PX focuses on constructs like fun [84, 257, 261, 274], immersion [7, 67, 127, 212], flow [87, 103, 316, 457–459], engagement [56, 353, 354, 519], and presence [44, 205, 462, 463, 527]. However, there is no agreement on which constructs are a part of PX, let alone their definitions, scopes, interactions, and relationship to various game elements [51]. Consider *immersion* as both a formal construct and colloquial term; a player can be "immersed" in the game environment (aesthetics), in the gameplay (mechanical), or

in the story (emotional). While game atmosphere [393] and character interactions [50] are shown to enhance PX, immersion models (e.g. 7, 67, 127, 212) are generally biased towards gameplay/challenge-based immersion. The qualities often associated with immersion (e.g. lack of time awareness, increased focus) are also commonly associated with flow, and even at times engagement. Common constructs, like flow, immersion, fun and engagement, also over-represent positive experiences, despite negative experiences (e.g. frustration, anxiety, tension) enhancing PX [212, 236, 467]. Overall, the constructs we have are inherently limiting the types of experiences we can discuss.

Despite unclear theory, many PX measurement tools (e.g. 4, 74, 110, 378, 403) capture these constructs. Recent work highlights the convergence of these tools [109], and their lack of validation (e.g. 109, 217, 228, 255). Without clear theory, it is hard to assess whether a tool's assumptions and measures are reasonable [51]. Consider widely used Player Experience Needs Satisfaction (PENS) [403], which applies Self-Determination Theory (SDT) to PX. We are unsure whether its constructs measure PX [486] or if SDT is suitable for all types of experiences [109, 377]. Tools looking to assess constructs thus inherit the issues of the constructs.

ExpTypes can be a useful bridge towards a general theory of PX. ExpTypes represent a bounded PX, defined by the terms of the player-game interactions for that context. This clarity of the system components means we can look for ways to measure each ExpType. We can leverage the clear boundaries and measurements of ExpTypes to compare experiences between games. ExpTypes integrate existing PX work through mapping elements to types; for example, Table 1.1 maps types of fun to their related ExpTypes. This way ExpTypes builds on existing PX work, while making clear the boundaries and limitations of individual work and concepts.

Kind of Fun	**Definition of Fun**	**Experience**
Sensation	*Game as sense-pleasure*	Mechanical, Aesthetic
Fantasy	*Game as make-believe*	Aesthetic, Emotional
Narrative	*Game as drama*	Emotional
Challenge	*Game as obstacle course*	Mechanical
Fellowship	*Game as social framework*	Emotional, Socio-cultural
Discovery	*Game as uncharted territory*	Emotional, Socio-cultural
Expression	*Game as self-discovery*	Emotional, Socio-cultural
Submission	*Game as pastime*	Mechanical, Aesthetic

Table 1.1: 8 Types of Fun [261] related to the 4 Experiential Types.

Our work furthers this idea of ExpTypes by exploring MX. We particularly focus on its two components: *mechanical achievability* (whether the gameplay is possible for the player to beat), and *mechanical difficulty* (how the gameplay's design affects the mechanical achievability). We start with MX because every game has one, and it overlaps with other PX constructs. In this way, MX is a barrier to the rest of PX since not being able to engage with a game's MX means a player cannot engage with the game's other ExpTypes.

1.1 Research Questions

We want to know how a player's cognitive and motor abilities interact with gameplay challenges to understand mechanical experience through mechanical achievability and difficulty. To do so, we ask:

RQ1: How are cognitive and motor abilities used to interact with various challenges?
RQ2: What are the effects of cognitive and motor overloading on the mechanical achievability of a challenge?
RQ3: How can designers use this knowledge?

1.2 Problem Statement

MX is the relationship between the player's abilities and the gameplay challenges. MX affects PX through playability. It determines playability through *mechanical achievability* (whether the gameplay is possible for the player to beat), and *mechanical difficulty* (how the gameplay's design affects the mechanical achievability). To model MX we need to model mechanical achievability and mechanical difficulty for different gameplay challenges.

We address RQ1 by finding the mechanical achievability for challenges. For each gameplay challenge we create a ***competency profile*** (set of cognitive and motor abilities that characterize a task [2]). Competency profiles let us compare the challenge's ability requirements to the player's abilities to determine if the game is possible to beat.

We explore RQ2 through the mechanical difficulty of a challenge and its relationship to mechanical achievability. We consider a challenge's main source of mechanical difficulty to be the ability with the highest proficiency requirement (i.e. its limiting ability). We adjust the load on the main source of mechanical difficulty by tweaking its associated game element, allowing us to test overload.

While we cannot perfectly address RQ3, we present a working representation of MX in the form of a jutsu (expanded from 441). A jutsu visualizes the MX of a particular challenge for a particular player profile. Jutsus can be incredibly powerful tools for designers to help...

- tailor game experiences,
- explore the effect of changes in a game's mechanical experience,
- create new kinds of challenges,
- and explore other avenues we have yet to see.

Overview of Scoping Decisions

We need to scope this work in a way that allows us to explore the most apparent elements of MX and highlights the potential merits of this idea. As we do not have specific existing work on MX for various gameplay challenges, we need to further refine the scope of our problem.

We believe the study of MX needs to start at constructing and validating preliminary competency profiles for challenges. To this end we would need to both create a hypothetical competency profile, and have some way to test that the relationship between its abilities and challenge success. We anticipate this could be difficult as challenges require many interacting

abilities which are themselves affected by potentially many design decisions. Therefore it seems prudent to scope our efforts to a challenge (or potentially set of challenges) that has a small set of required abilities, few gameplay mechanics (ergo design decisions) and an easily observable/apparent limiting ability that connects to those mechanics.

Scoping to button mashing challenges. For simplicity of generating competency profiles that can be easily validated we scope our work to studying button mashing challenges. These are mechanically simple, motor-focused gameplay challenges that have a clear limiting ability. We elaborate on this decision in Ch. 13.3.

1.2.1 Specific Problem

To summarize, **we aim to conduct a preliminary exploration of the mechanical experience idea by exploring the mechanical achievability and mechanical difficulty of button mashing challenges as it relates to their limiting ability**. We do this by constructing competency profiles for each button mashing challenge, and comparing them to a player profile in a jutsu. **Our specific problem is creating and validating competency profiles for button mashing challenges.** To accomplish this, we will:

- construct a list of human cognitive and motor abilities (Part II),
- design and validate a game-based player profiling tool (Ch. 11),
- define a set of atomic button mashing challenges (Part III),
- validate our preliminary competency profiles for our button mashing challenges (Ch. 16),
- explore overload's effects on our button mashing challenge competency profiles (Chapter 17), and
- construct jutsus for our button mashing challenges which reflect how changes to the challenge design affect the ability requirements of the gameplay, particularly around the limiting ability (Part IV),

1.3 Guide to the Thesis

1.3.1 Thesis Structure

We break down this thesis into parts which focus on building towards answering our research questions and solving our problem statement. Each part is self-contained with its own literature review relevant to its topic, and conclusion which summarizes the part's important points and deliverables.

Part I explores game-related experiences, including PX. It culminates in the requirements and design decisions behind our various models, and serves as guiding direction to the rest of the thesis. Part II presents the construction, and validation of the Player Model. Part III expands the Challenge Model and empirically validates the competency profiles. We also explore the effects of ability loading on competency profiles and perceived experience in this part. We end our thesis with Part IV, which puts together the Player and Challenge models into a clear representation of MX for our challenges (i.e. jutsus).

1.3.2 Novelty

While we presented the basics of jutsus in Soraine and Carette [441], we improve on the idea through *expansion* and *validation*.

Expansion. We improve the existing player model by adding more detail to the cognitive abilities. This makes jutsu more robust as they can interpret more complex challenges with the inclusion of cognitive abilities. We use the expanded player model to revise the original challenge models. We perform close readings of their gameplay to identify underlying challenges and their approximate competency profiles. This lays the groundwork for exploring more complex challenges that are the combination of many small challenges.

Validation. We create player profiles from empirical data, and experimentally validate the challenge competency profiles. These experiments make the jutsu comparing player and challenge profiles more reliable as it is based on quantifiable data. We also empirically explore the effects of changing the loading of abilities in the competency profile to get a complete picture of the relationship between player and game.

1.3.3 Contributions

In the process of answering these questions, we create many artifacts that summarize our understanding of different concepts and methods. As well, we produce results from our two main studies that aim to add concrete evidence for our ideas. We separate our contributions into conceptual, methodological, and experimental.

Conceptual contributions: These are the theoretical frameworks, and concepts that we created for the thesis. These contributions are the result of synthesizing existing literature or other information throughout the work. As such, while they are novel creations of our thesis, they may not be fully explored or validated due to scoping limitations in our work.

- the presentation of the Experiential Tetrad (i.e. ExperT, Fig 3.1),
- a definition for mechanical experience (Def 3.1),
- the definition of player profiles (Def 3.2),
- the definition of challenge competency profiles (Def 3.3),
- the definition of mechanical achievability (Def 3.4),
- a model of human motor (Ch. 7) and cognitive (Ch. 8) abilities for player and competency profiles, and
- a proposed outline for a Jutsu Framework (Ch. 21.3.4).

Methodology Contributions: These are the methods we create over the thesis to produce hypothetical and real player profiles and competency profiles. Included in these methods are ways to visualize and interpret the profiles. These contributions are closely tied to the experimental contributions that test the methods and validate the profiles.

- a visualization method for player profiles (Ch. 12),
- a method for finding challenge competency profiles (Ch. 13.2),

- a visualization method for challenge descriptions and competency profiles (Ch. 19),
- a visualization method for mechanical experience resulting in the creation of three jutsus (Ch. 20).

Experimental Contributions: These are the two major studies we conduct for our thesis. They are exploratory in nature, and produce preliminary evidence for our conceptual contributions. Through these studies we aim to validate our methodology contributions and chip away at answering our research questions.

- an empirical validation of a measurement battery for player abilities (Ch. 11),
- a study addressing our research questions, sub-divided into three specific studies:
  - an empirical validation of three button mashing competency profiles (Ch. 16),
  - a study on the effects of overload/underload of the limiting ability on mechanical achievability (Ch. 17), and
  - a mixed-methods study on player experience of challenges at different difficulties/limiting ability loads (Ch. 18).

1.3.4 Speculating on Potential Impacts of Contributions

We believe the concepts of MX, competency profiles, and jutsus are broadly useful to games user research (GUR) and games studies (GS), with potential effects in gamification and other areas of game-related studies. Individually each specific contribution creates a base from which future research can expand. **ExperT**'s separation of experiences sets up future work to explore what challenges and achievability look like for different ExpTypes. Our **profiling tool** adds to GUR's growing body of work on stealth assessments, and opens up work for integrating player modeling with dynamic difficulty methods. The button mashing **challenge models** and **jutsu** are a starting point for further investigations into the relationship between individual mechanics and game difficulty, as well as for assessing the effect of different controllers on experience. The whole thesis also serves as a methodology for conducting this kind of broad modeling work.

We take a minute to highlight the two contributions we think are most important from the thesis: ExperT and the Jutsu-modeling approach. The following are our speculations about the impact these contributions could have on the larger domain.

ExperT: A Theoretical Framework. ExperT highlights boundaries for and connections between ExpTypes. Adopting ExperT as a theoretical framework would make it easier for experience-focused research to be clear about their context and limitations. It would also allow easier integration or comparison of experience-focused research between GS and GUR.

Jutsu-modeling Approach. Under ExperT, our work acts as an outline for modeling an individual experience in a measurable way. Exploring other ExpTypes could take a similar approach of defining a Player and Challenge model, with the main focus being on defining the appropriate set of skills needed to interact with that ExpType. For example, an emotional experience (EE) derived from characters and narrative may require skills like recognizing

emotions in characters, or understanding story structures. With a set of skills player's need to engage with the experience, modeling then takes the same approach as our work.

Part I
Game-Related Experiences

Here we explore existing work on game-related experiences to flesh out our understanding of mechanical experiences. We start by reviewing work on player experience, focusing on holistic models, constructs, and measurement tools (Ch. 2). From there we explain the Experiential Tetrad, our theoretical framework of experience types and their relationships, and mechanical experience (Ch. 3). We also explain how the Experiential Tetrad integrates existing experience-related work from different domains, and expands into non-player experiences. We synthesize the literature and our ideas of the Experiential Tetrad into requirements and design decisions for modeling mechanical experience, and formal definitions for key components of our model (Ch. 4).

Chapter 2

Player Experience

To understand mechanical experience (MX), we need to first look at player experience (PX).

PX does not have a singular definition [51, 319], but it is generally understood to be a multi-dimensional [4, 200, 204, 403, 415], context-dependent [319] phenomena. PX colloquially describes everything from a game's intrinsic qualities like accessibility and difficulty, to extrinsic qualities, like an individual's opinion of the game. PX is a competing term against *game experience*, *gaming experience*, and *user experience*, which focus on different parts of the experiential system (i.e. player-game relationship) [51]. We incorporate all of these terms into our search for existing PX work, and consider them equally integral parts of the tapestry for understanding game-related experiences.

We use this chapter to summarize our literature review into PX and game-related experiences. This is non-comprehensive, as we look to present a sample of works that illustrate concepts in PX. We group our curated examples into high-level frameworks, PX constructs, and PX measurement tools. We do not expect existing work to discuss PX through experiential types (ExpTypes), let alone specifically through MX as the relationship between human abilities and gameplay challenges. However, we aim to use this survey as a guide post for our conceptualization of PX. We distill from this search design considerations that our own theoretical framework should incorporate.

2.1 Models and Frameworks

Game design is generally understood and taught as the creation of "good" experiences [51, 107]. In response, many frameworks exist to guide designers through game design and analysis in hopes of crafting "good" PX [358]. At this point we focus on frameworks that describe the relationship between player and game, ignoring for the moment those that focus exclusively on the player (e.g. Bartle's Types [30], Yee's Motivations [534])[1]. We organize our coverage into *game-focused frameworks* which model the game as components the player interacts with, and *experience-focused frameworks* which try to explore some of the phenomenological elements.

[1] See the Player Profiling (Ch. 6)

2.1.1 Game-Focused

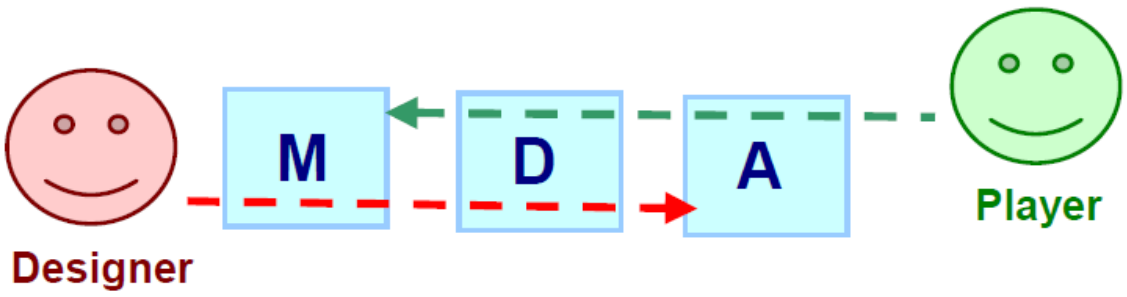


Figure 2.1: Mechanics, Dynamics, Aesthetics Framework as presented in [200].

The **Mechanics, Dynamics, Aesthetics (MDA) Framework** [200] (Fig. 2.1) describes games as a medium of communication between designer and player. The game exists at three layers: mechanics ("the components of the game at the level of data representations and algorithms"), dynamics ("the run-time behaviour of mechanics acting on player inputs and each other's outputs over time"), and aesthetics ("the desirable emotional responses evoked in the player"). The layers are organized to represent the actor's control over the system; designers have most control over the mechanics, while players have control over the aesthetics (i.e. PX). As noted by many (e.g. 113, 114, 512, 525), MDA has a host of issues, like:

- the meaning of the layers are unclear;
- game mechanics are overly emphasized compared to other design elements;
- designers are shown to have little-to-no control over the PX;
- it focuses on games for entertainment, and is not applicable to serious games or gamified experiences.

There have been many attempts to address these problems through extended or new frameworks (e.g. 113, 114, 512, 525). For example, the **Design, Dynamics, Experience (DDE)** framework [512] attempts to address all the above issues through reframing and expanding each layer. Notably, DDE integrates the player's personal context into the experience process through the *player-subject* (a mental persona that the player inhabits when engaging with the game [434]). However, revisions and expansions of MDA have not received much traction in the game design or research community.

Schell's **Elemental Tetrad** [415] (Fig 2.2) models games as four basic elements: mechanics ("the procedures and rules of the game"), story ("sequence of events that unfold in the game"), aesthetics ("how the game looks, sounds, smells, tastes, and feels"), and technology ("any materials and interactions that make your game possible"). These elements are meant to be equally important to the game experience, with connections between them showing how they must support each other. They are visually organized by how aware the player is of each aspect during play (i.e. their visibility). The Elemental Tetrad considers games as a medium for experiences, but is not clear on how its components come together to *create* an

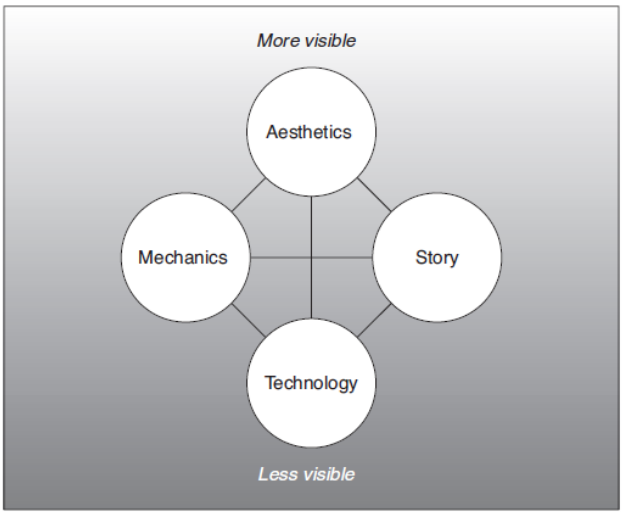


Figure 2.2: Elemental Tetrad as presented in [415].

experience. As a descriptive framework it also suffers from lack of clarity about what boundaries exist between the components, and lack of generalizability to non-traditional games, such as those that tell story through emergent narratives or procedural mechanics.

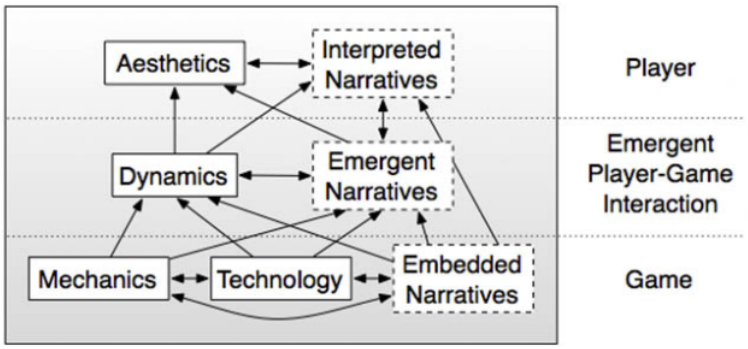


Figure 2.3: MTDA+N as presented in [386].

The **MTDA+N**[2] framework [386] address these concerns by combining the mechanics, dynamics, aesthetics of MDA, with the technology and story (renamed narrative) from Elemental Tetrad (Fig 2.3). It outlines three different types of game narratives: embedded (the plot), emergent (perceived personal story created from player interactions), and interpreted (the player's representation of the embedded and emergent narratives). MTDA+N organizes itself based on which actor (player or game) controls the component; therefore the game controls the mechanics, technology and embedded narratives while the player controls the aesthetics and interpreted narratives. While it attempts to be more specific, MTDA+N inherits the specificity and scope problems of MDA and the Elemental Tetrad.

Gibson's **Layered Tetrad** [157] describes PX as the Elemental Tetrad interpreted through inscribed (game on its own), dynamic (game during play), and cultural (game in relation to

[2]**M**echanics, **T**echnology, **D**ynamics, **A**esthetics plus **N**arrative.

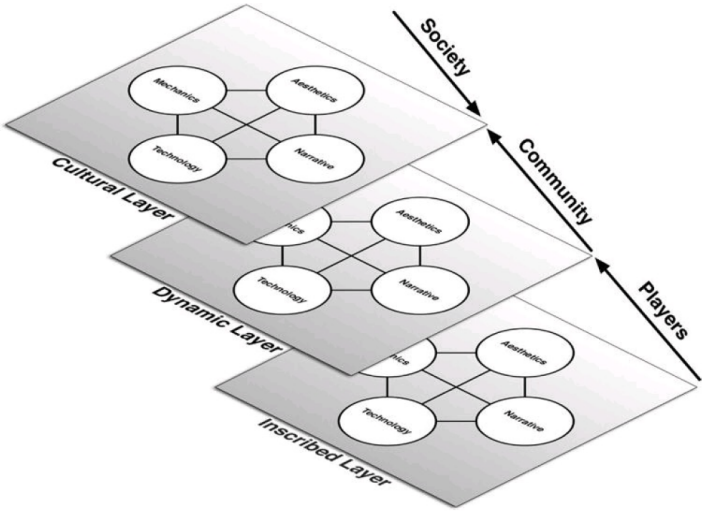


Figure 2.4: Layered Tetrad as presented in [157].

community) layers (Fig 2.4). The meaning of game components changes across the layers. For example mechanics are the rules of the game at the inscribed layer, the "procedures, strategies, emergent game behaviour and eventually the outcome of the game" at the dynamic layer, and elements like game modding and the impact that the emergent dynamic gameplay had on larger society at the cultural layer. The changing meaning across layers makes this framework more generalizable, but boundaries between components and their individual meanings can still be unclear. For example, in the dynamic layer aesthetics covers everything from "procedural art to the physical strain that can result from having to mash a button repeatedly", which makes it hard to separate from both technology and mechanics.

2.1.2 Experience Focused

Nacke, Drachen, and Göbel [319] propose that PX exists as three related gameplay experiences: the game system experience, the individual PX, and a framed context experience. Nacke and Drachen [315] further this understanding by synthesizing empirical and theoretical work into a **PX framework** (Fig. 2.5). The PX framework shows how these experiences exist at different levels of abstraction (game system as concrete, and framed context as abstract) and are impacted by time. The purpose of this framework was to help situate existing and future work so there could be more awareness around boundaries and proper methodologies.

2.1.3 Summarizing Frameworks

The frameworks are predominantly descriptive, biased towards particular styles or types of games (hence not generalizable), unclear in the definitions and scopes of their components, and often do not situate the experience in a specific context (except for Layered Tetrad, and PX Framework). While they do not capture everything we want in a framework we do think they provide valuable insight into the multi-dimensional nature of PX, the ways that PX

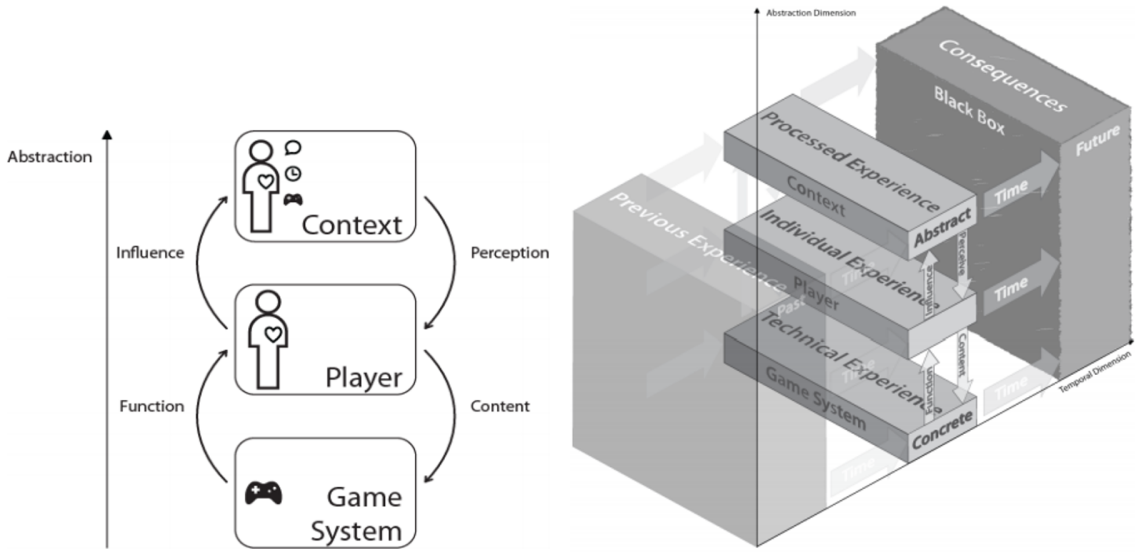


Figure 2.5: Player experience framework and levels of abstraction as presented in [315]

cannot be separated from the game elements.

2.2 Constructs of the Player Experience

Constructs scope PX down to specific universal and measurable experiences. While significant work has been done defining and measuring constructs, there is no consensus on which are inherent to PX. Borges et al. [51] found 70 different constructs studied by researchers, of which only 22 appeared more than once and in those there was significant overlap in their definition and scope. In this review we discuss fun, flow, immersion, presence, and engagement as they are frequently studied constructs.

2.2.1 Fun

Malone [274] and Malone [275] views fun as intrinsic motivation inspired by challenge, fantasy, and curiosity. Challenge is created through providing goals with uncertain outcomes; it engages the player's self esteem. Fantasy is a sensory element which engages the player's emotional needs. Curiosity is divided into two types: sensory and cognitive. Both serve as vehicles for continued engagement. Malone's view of fun is limited in scope to single player games, and by its context in motivation for children's education. While Malone [275] describes what games should have, they do not explain how to create the right conditions, or to measure fun (except for challenge).

LeBlanc [261] outline a taxonomy of fun[3]:

1. sensory,
2. fantasy,
3. narrative,

[3]This was also presented in their MDA paper [200]

4. challenge,
5. fellowship,
6. discovery,
7. expression, and
8. submission.

These experiential categories are broad, and treated as self explanatory, but there are questions of boundaries (e.g. "what's the distinction between Fantasy and Narrative?" [405]) and interactions between categories (e.g. "doesn't self discovery occur in a challenge?" [405]).

Lazzaro [257] describes four "fun keys" (i.e. desired play styles) based on player's facial expressions when interacting with game mechanics:

1. easy fun ("inspiring imagination, exploration, and role play"),
2. hard fun ("challenge and mastery"),
3. serious fun ("changing a player's internal state or doing real work"), and
4. people fun ("social interaction").

Lazzaro highlights how different mechanics can be used to guide players towards these different play styles/experiences (Fig. 2.6). Players regulate their overall experience by cycling through the different play styles.

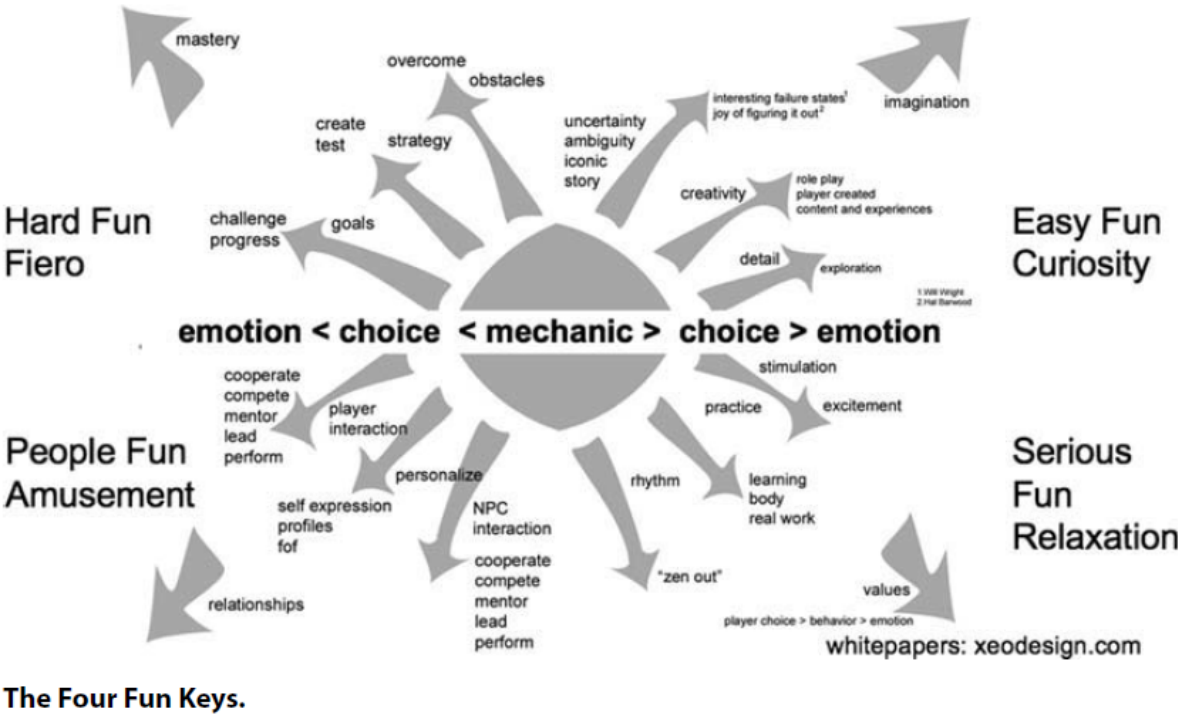


Figure 2.6: Fun Play Styles as presented in [257].

Summary. Fun is highly coupled with discussions about "good" PX in the larger gaming culture (e.g 84) leading to ample ontological work to study it. Still it is poorly defined, and

difficult to measure or create. Outside of issues with measurement fun as a concept is too linked to positive experiences, and so limits the type of experiences that can be explored (e.g. 7, 200).

2.2.2 Flow and GameFlow

Flow [103] is an emotional state of total absorption into an autotelic task (i.e. a task that was *intrinsically motivating* and offered few to no conventional rewards). Flow requires nine specific conditions be met:

- There are clear goals every step of the way,
- There is immediate feedback to one's actions,
- There is a balance between challenges and skill,
- Action and awareness are merged,
- Distractions are excluded from consciousness,
- There is no worry of failure,
- Self-consciousness disappears,
- The sense of time becomes distorted,
- The activity becomes autotelic.

However, the common takeaway is that Flow is achieved in games when a player's skill matches the level of difficulty in a task (see Fig. 2.7). This focus on challenge-based experiences means it is not well suited to other types of PX and non-challenge-based games.

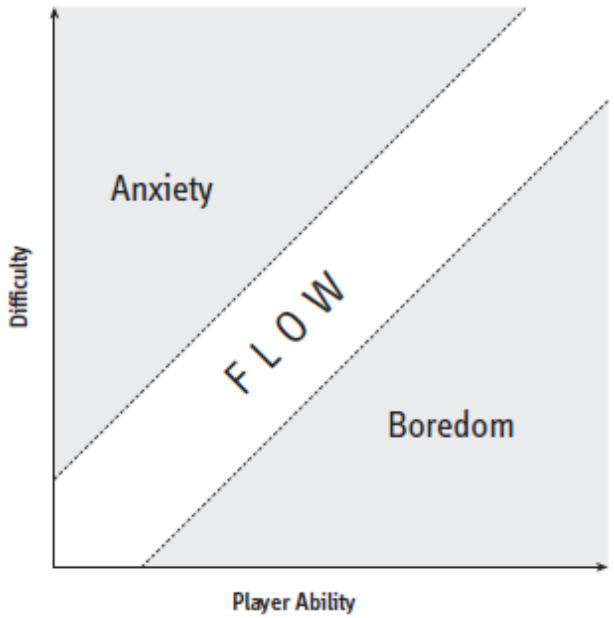


Figure 2.7: Flow state as a relationship between player abilities and difficulty as described by Csikszentmihalyi [103], picture reproduced from [7].

GameFlow [457] adapts Flow to games, defining an enjoyable game as requiring concentration, challenge, player skills, control, clear goals, feedback, immersion and social interaction. However, these elements are highly genre dependent; for example, feeling a sense of ***control*** and impact on the game world is something more relevant in role playing games than strategy games. Sweetser et al. [458] argues all genres would improve by meeting these GameFlow criteria. But, their argument is based on comparing GameFlow and Metacritic scores for first-person shooter and adventure games. These genres are biased towards the criteria (albeit in different ways), and the choice of comparing expert reviews to general player reviews is questionable. Expert reviews tend to be more balanced and favourable as opposed to player reviews (which are more polarized); experts consider game meta-properties and focus on descriptions of game elements over gameplay experiences [409]. Outside of genre-bias, GameFlow relies on poorly defined sub-dimensions like immersion, whose criteria is a list of qualities of Flow (altered senses of time, merging of actions and awareness, and less awareness of "distractions"/their surroundings) and overlaps with other dimensions like concentration. While this is somewhat addressed in Sweetser, Johnson, and Wyeth [459] it does not resolve all the underlying problems.

Summary. Flow has been applied to games many times (e.g. 87, 316) because it aligns closely with challenge-based PX and provides seemingly actionable design guidelines. However it is limited in the types of games it can apply to, and the types of experiences it can describe. Despite this, it continues to be a commonly discussed construct and is incorporated into many others.

2.2.3 Immersion

Adams [7] explores tactical, strategic, and narrative immersion. *Tactical immersion* (aka "Tetris Trance") describes "being in the groove"; it is basically Flow, where the player is mechanically focused and their skills align with the challenge. *Strategic immersion* describes cognitive involvement in the game via total concentration in planning, observing, and calculating moves. *Narrative immersion* is engagement with the story, world, and characters. This is a non-comprehensive, descriptive model that highlights a mechanical/challenge-based skew to immersion.

Brown and Cairns [67] propose immersion is a process of moving from engagement, to engrossment, and finally total immersion. Engagement is tightly coupled with accessibility and gamer investment; barriers to engagement include gamer preferences, control and feedback accessibility, and the time they put into the game. Engrossment is when the game begins to directly influence the player's emotions; they describe this as being less aware of their surroundings and becoming absorbed in the visuals, plot and gameplay (like Flow). Total immersion describes feeling Presence in the game; its barriers are empathy and atmosphere. Generally this model overlaps in areas with the concept of Flow. The distinction between each phase is fuzzy, and their examples are very sensory focused – both of which make it difficult to figure out applying to game design.

Jennett et al. [212] model immersion with three player factors (cognitive involvement, real world dissociation, and emotional involvement) and two gameplay factors (challenge, control). While their findings on this work were not significant, they believe the factors

support Brown and Cairns [67] view of emotional involvement (i.e. engrossment) as important to immersion.

Ermi and Mäyrä [127] present three types of immersion: **S**ensory (i.e. audio-visual), **C**hallenge-based (i.e. player skills vs. gameplay), and **I**maginative (i.e. narrative, world, characters) (**SCI**). They propose immersion is one part of a game experience, created from the system-level interaction between game and player, situated in social context (Fig. 2.8). While the elements that contribute to these immersions seem more distinct, it is unclear how much they affect each other and overlap.

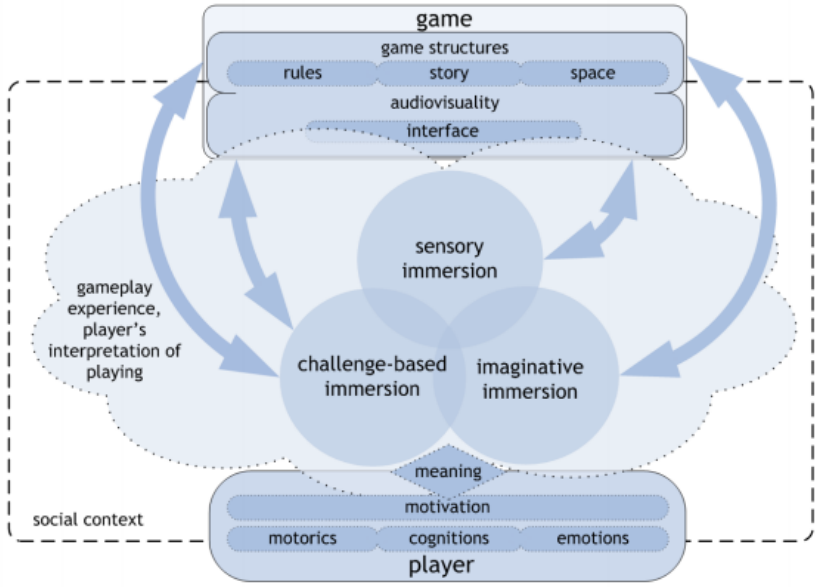


Figure 2.8: Sensory-Challenge-Imaginative model as presented in [127].

Summary. Colloquially immersion is understood as "good" PX. Players and designers alike intuitively know when they are in it, but have a hard time consistently reconstructing its characteristics. Immersion's scope overlaps with other constructs like Flow and Presence, and so biases towards challenge-based and sensory-based experiences.

2.2.4 Presence

Tamborini and Skalski [463] explore three types of presence: spatial, social, and self. *Spatial presence* is the sensory experience of physically inhabiting the virtual space, made possible through immersion, and involvement [462]. Spatial presence is affected by the game's interactivity (the ability to influence the form and content of the virtual environment) and vividness (the ability to produce a rich sensory environment). *Social presence* is the emotional experience from perceiving virtual actors (e.g. NPCs) as social actors (i.e. real people). Social presence has three dimensions [44]:

- copresence (awareness of spatial presence of another, sometimes including mutual awareness),

- psychological involvement (a sense of intelligence in the other agents), and
- behavioural engagement (participation in social behaviours).

Self-presence is when the player experiences their virtual avatar as if it were themselves. The types of presence overlap, leading some to think that self-presence is just a part of spatial presence (e.g. 205, 527). It is also unclear how specific game elements affect the different types of presence. Intuitively avatar control and better representation should greatly impact self-presence, but a recent meta-analysis found avatar customization/selection only resulted in small impacts on presence, and no significant impact on self-presence [83].

Summary. Presence is about feeling "in the game". Like Immersion, players feel presence but cannot conceptually separate or articulate it from other constructs. Internally its dimensions overlap significantly, and its relationship to specific design choices is unclear.

2.2.5 Engagement

O'Brien and Toms [354] characterise engagement as nine attributes (challenge, aesthetic and sensory appeal, feedback, novelty, interactivity, perceived control and time, awareness, motivation, interest, and affect) that ebb and flow in intensity during play[4]. O'Brien and Toms [353] refines this into six factors (focused attention, perceived usability, endurability, novelty, aesthetics, and felt involvement) which were not fully supported in exploratory factor analysis (endurability, novelty and felt involvement merged into a larger satisfaction construct) [519].

Summary. Engagement describes investment into a game; players need to be engaged in order to access flow, and immersion (though it can co-occur with presence) [56]. Engagement research is split between viewing it as an experiential construct or a motivation for play, with more work falling on the motivation side [56]. Therefore work addressing engagement as a construct is less theory-based, and more measurement focused (e.g. 66, 403, 519) so will be covered in the next section.

2.2.6 Summarizing Constructs

Constructs are great at speaking to common gameplay experiences, but struggle to be precise in scope, definitions, and relationship to game designs. They frequently overlap along challenge-based biases (i.e. Flow, challenge-based Immersions) or sensory-biases (i.e. Presence, sensory-based Immersions). Constructs also prioritize positive, optimal experiences; contemporary research is starting to push back on this focus by highlighting ordinary PX [485], and how negative experiences like frustration, anxiety, and tension enhance PX [212, 236, 467]. Overall, constructs provide insight into representing common experiences without overrepresenting positive or optimal PX, and the importance of clear scopes, connections to design elements, and explanations on how dimensions interact.

[4]They used the threads of experience framework [287] to map the attributes to a particular stage of experience

2.3 Measurement Tools

Measurement tools capture the holistic PX during or right after a play session. These tools may have their own PX models, implement a set of constructs, or apply theories from other domains to games. These tools inherit issues of overlapping scopes and assumptions from their theoretical roots, causing many to converge [109]. We review the Core Elements of the Gaming Experience (CEGE), the Video Game Demand Scale (VGDS), Game Experience Questionnaire (GEQ), Game Engagement Questionnaire (GEngQ), the Challenge Originating from Recent Gameplay Interactions (CORGIS), Player Experience Needs Satisfaction (PENS), the Player Experience Inventory (PXI), and Ubisoft Perceived Experience Questionnaire (UPEQ). We organize this section based on the theoretical background of the tools.

2.3.1 Independent Models

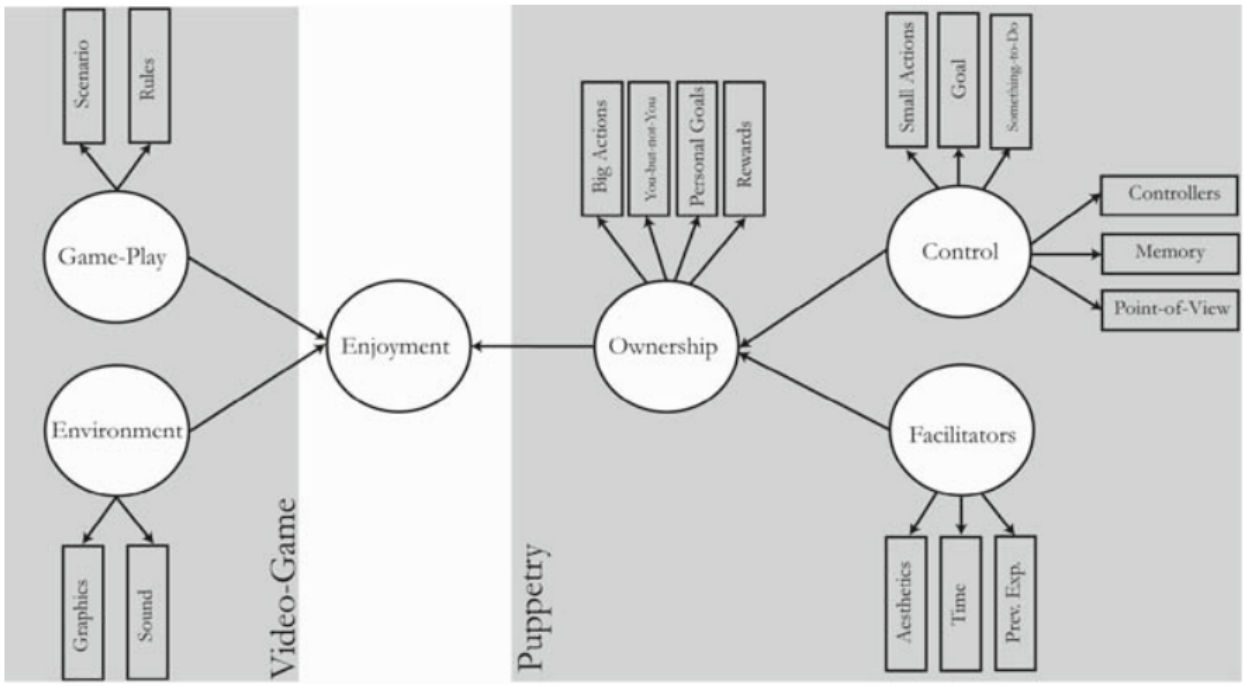


Figure 2.9: Core Elements of the Gaming Experience [74].

CEGE [74] has two "core elements": the video game and puppetry (the interaction between the game and the player). The video game is made up of the gameplay (*rules and scenarios*) and the environment (*graphics and sound*). Puppetry is made of control (*goals, basic controls, controllers, memory, something-to-do, and point-of-view*), ownership (*big actions , you-but-not-you actions, personal goals, and rewards*), and facilitators (*time, aesthetic values, and previous experiences with similar games*). Control represents the in-game actions and events; ownership reflects the player agency; facilitators make it easier for players to feel control and ownership in a game. The observable elements (square boxes Fig. 2.9) are

measured via questionnaire, and their combination defines the latent (unobservable) properties. CEGE proposes the presence of these elements allow (but does not guarantee) positive PX; while their absence ensures a negative PX. They derive these attributes from coding game reviews, and interviewing five individuals (one game designer, two game reviewers, and two players). There is inherent bias in basing most ideas off reviews considering they are primarily a sales tool reflecting expert opinions (as previously stated).

VGDS [55] is a 26-item scale measuring the cognitive, emotional, social, and physical requirements exerted by the game on the player. *Cognitive demand* is the engagement of mental (predominantly attentional) faculties by the game. *Emotional demand* focuses on player-centred (e.g. emotional investment, emotional responses) and context-centred emotions, predominantly via game narrative. *Social demand* measures game-induced and player-initiated sociality that could be applied to both NPCs or other players.*Physical demand* reflects effort put into the game as either *controller demand* or *exertion*. VGDS seems to measure player perceptions of gameplay as opposed to their actual abilities related to gameplay [140], which reflects how it was validated via player recall of a favoured game experience.

2.3.2 Implements Constructs

GEQ [378] is three widely used, but unvalidated [217, 255], modules:

Core GEQ: measures *sensory and imaginative immersion*, *tension*, *competence*, *flow*, *negative and positive affect*, and *challenge* while gaming;
Post-game (PGQ): measures experience (unknown factors) right after the gaming session; and,
Social presence (SPGQ): measures *empathy*, *negative feelings*, and *behavioural involvement* with co-players.

The method for generating questions seems open to inherent bias from the focus group, which could affect GEQ's ability to capture nuanced differences in PX for different games. Consider two players looking for a relaxing experience after a busy day. One chooses to play League of Legends [395], the other plays Animal Crossing [342] — games with very different core gameplay loops, and intended experiences. Both end up reporting the same relaxing experience about the game in their responses to the questionnaire (since questions involved are high level such as "I enjoyed it" and "I felt successful"). Therefore, differences in player preferences and gameplay design get subsumed in the larger subjective feeling, making it difficult to trace back why the PX was good or bad.

GEngQ [66] is a 19-item questionnaire which measures a player's potential for becoming engaged during gameplay. GEngQ views engagement as distinct levels, specifically: immersion, presence, flow, and psychological absorption. We cannnot find external support to view these constructs as linear levels of engagement, and are unclear why the levels are ordered this way. GEngQ builds on existing measures for immersion, presence, flow, absorption and dissociation by refining the items via two focus groups (one with children and one with adults). It is unclear how GEngQ differentiates between Presence items like "things seem to

happen automatically" and Flow items like "playing seems automatic". GEngQ also highly correlates with the Immersive Experience Questionnaire (IEQ), which could indicate they are actually measuring the same construct [109].

CORGIS [110] measures *perceived challenge* (effectively the PX of player-game interactions at a particular skill level) through a 30-item questionnaire. They identify four challenges: physical (e.g. speed, reaction time, endurance, accuracy), cognitive (e.g. memory, problem solving, reasoning), decision-making (e.g. making potentially regrettable decisions), emotional (e.g. emotional connection, narrative tension). Through comparison study with VGDS, CORGIS is implied to capture player opinions about their abilities in the game [140]. As a newer tool, there has not been significant work done to further validate or expand CORGIS.

2.3.3 Theory-based

PENS [403] explores PX through self-determination theory (SDT) using four subscales: in-game competence, in-game autonomy, relatedness, presence, intuitive controls. Competence focuses on perception of challenge and difficulty. Autonomy is the perception of opportunities in the game. Relatedness is how the player feels about their interactions with other players. Presence covers physical, emotional and narrative immersion. Intuitive controls measures the experience with the interface. While widely used, we are unsure whether PENS constructs actually measure PX [486], if its factor model is valid [109, 217] or if its underlying theory is suitable for all types of experiences [109, 377].

PXI [4] implements Means-End Theory[5] to measure PX as functional and psychosocial consequences through a 30-item questionnaire. Functional consequences are the "immediate, tangible consequences experienced as a direct result of game design choices". Psychosocial consequences are "the emotional experiences, as a second order response to the game design choices." These consequences are measured by attributes (summarized Table 2.1).

Functional	Psychosocial
ease of control	mastery
progress feedback	curiosity
audiovisual appeal	immersion
goals and rules	autonomy
challenge	meaning

Table 2.1: Attributes for Functional and Psychosocial Consequences

UPEQ [20] is an industry-driven, game-focused survey, measuring Autonomy, Comeptence, and Relatedness from SDT. UPEQ focuses on these constructs *because* of their focus on positive PX. Like PENS, UPEQ inherits the basic criticisms of SDT as applied to games. UPEQ's factor model was not supported in independent study using non-Ubisoft games [228].

[5]The idea that user product preferences are a function of whether the product attributes (means) are likely to produce results (consequences) that align with their values (ends).

2.3.4 Summarizing Measurement Tools

Measurement tools are highly specific, quantitative understandings of PX, which makes them attractive to research and industry. However, they are limited by their underlying theories (e.g. construct overlap, specific genre biases, emphasis on optimal and positive PX). As well, a tool's effectiveness is highly dependent on the context (temporal and experiential) in which it is administered. Most of the tools covered are proposed as being for directly after a gaming session. However, the validation studies frequently ask participants to recall a recent gaming experience, or often a memorable one (often interpreted as positive, good, or "favourite"). Reviewing these measurement tools highlights the importance of situating the experience, as well as the inherent desire and usefulness for measurable and actionable understandings of PX. We also particularly note VGDS and CORGIS as approaching our understanding of MX.

2.4 Lessons from Player Experience

Reviewing this literature gave us insight into our theory-crafting, and what kind of information we should scope into our problem or stay away from. We summarize these insights into design goals for any theoretical framework of PX we make:

- G1: must reflect the multi-dimensional nature of experience;
- G2: must clearly define the dimensions and their respective scopes;
- G3: dimensions should clearly relate to game elements and user interpretations;
- G4: must reflect the interconnected nature of experience;
- G5: should try to make it clear how the dimensions interact;
- G6: must capture the various contexts of experience;
- G7: should allow an unbiased description of various experiences (e.g. positive and negative, optimal and normal);
- G8: should provide a way to make apparent the position of the experience (e.g. player, spectator);

So while we do not use these works, their ideas inform our understanding of game-related experiences and MX, as we venture into creating our own theory in Ch. 3.

Take home points

From this chapter we learned the following meta-lessons:

- PX is many different connected types of experience, situated in specific context.
- Constructs are good at describing common experiences, but have difficulty clearly modeling them so we can systematically implement in games.
- PX measurement tools inherit the scoping and definition problems from underlying theory.
- A unifying PX framework that provides measurable and actionable information is desirable.

Chapter 3

Mechanical Experience of ExperT

We synthesize our understanding of player experiences (PX) and game-related experiences (GX) into a single theoretical framework, the ***Experiential Tetrad*** (ExperT, Fig. 3.1). ExperT outlines four base experiential types (ExpTypes): mechanical, emotional, aesthetic, and socio-cultural. These temporally situated types interact with and influence each other to create a holistic GX. We present two analytical lenses with ExperT, the player (pink) and spectator (green), which represent different types of relationships a person may have with a game and how those contextualize their GX. We use ExperT to frame our understanding of mechanical experience (MX), and elaborate on its components, mechanical achievability and mechanical difficulty.

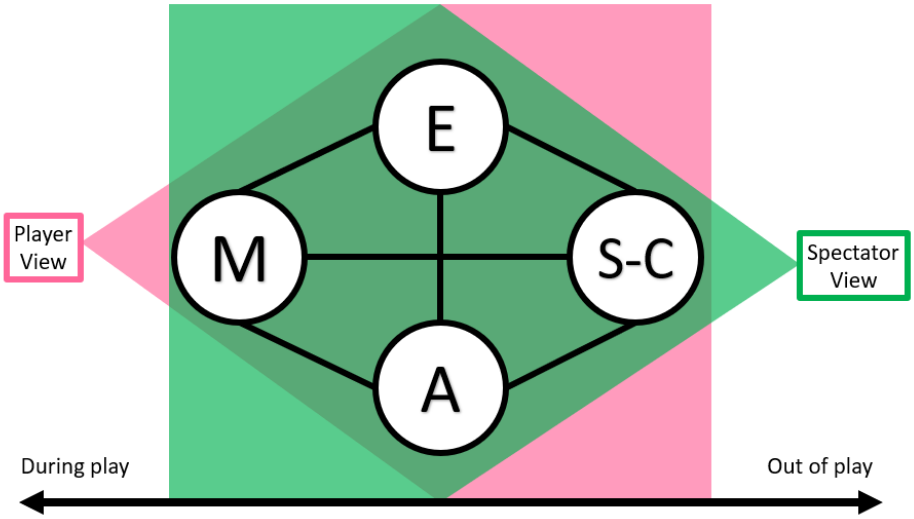


Figure 3.1: The Experiential Tetrad (ExperT). A framework of experience types, their relation, scopes and analysis lenses. **M**echanical, **E**motional, **A**esthetic, and **S**ocio-**C**ultural.

We start with explaining ExperT, its construction, ExpTypes, and analytical lenses. We present two examples that highlight use cases for ExperT. We then scope our work to focus on MX in the player lens and define its essential components. We also outline existing work that alludes to MX components. We end with a short summary and statement about next steps.

3.1 The ExperT Review

Our goal for ExperT is to integrate existing GX and PX concepts under a unified framework that allows for at minimum:

- clear delineation between types of experiences grounded in the game-human relationship,
- modeling of experiences and their interaction at different levels of abstraction, and
- equitable treatment of experiences (i.e. not favouring positive or optimal).

We draw inspiration for our process and presentation of ExperT from other conceptual frameworks like Aarseth and Grabarczyk's meta-ontology of games [3], Schell's Elemental Tetrad [415], and Wright, McCarthy, and Meekison's experiential framework [532].

3.1.1 ExperT Methods

We construct ExperT through a reflexive and iterative process of concurrent knowledge gathering and synthesis. We recognize that the curation of literature and resulting framework are reflections of our positionality as software engineering (SE) and human-computer interaction (HCI) researchers. We compensate by expanding our search outside our typical domain. However, certain understandings and design decisions for ExperT are inherently biased by our academic situation.

Knowledge Gathering. We first conduct a (non-comprehensive) narrative review of PX and GX work to understand how it shapes game design and game research (see Ch. 2). We cast a wide net, starting with general game design textbooks (e.g. 7, 58, 147, 157, 415) and industry-focused resources (such as GameDeveloper.com[1], and Game Developers Conference talks[2]), to identify concepts that transcend academia like game design frameworks (e.g. 200, 415), player typologies (e.g. 30, 534), experience taxonomies (e.g. 257, 261), and constructs like immersion, flow, and fun. We then search the ACM Digital Library, IEEE Xplore, and Google Scholar (for grey literature outside of technically-geared venues) for *player experience* and *game experience* broadly, and then again for specific concepts from the general resources. We focus on literature exploring or applying concepts, and use citation searching to find related work to enrich our understanding. We draw connections between concepts by comparing their goals, related game elements, similarities/differences, underlying assumptions and limitations. We continue this search and analysis process until our understanding of the work stabilized (i.e. new work did not change our underlying conceptual model of the theory or significantly innovate in its measurement or impact).

Knowledge Synthesis. We use our stabilized understanding of these concepts and existing work to derive design considerations for a unifying framework (see Ch. 2.4). We construct the base ExpTypes by grouping seemingly overlapping concepts, refining and analysing them as new information was found in our knowledge gathering. We summarize ExpTypes in Table 3.1 along with the related game elements and concept keywords.

[1] Formerly Gamasutra
[2] via YouTube and the GDC Vault

ExpType	Resulting from...	Related Concepts (Keywords)
Mechanical	Gameplay (e.g. mechanics, controls)	Mechanics, challenge, tactics, strategy, difficulty, hard, flow, control, competence, usability, and interactivity
Aesthetic	Art Direction (e.g. audio-visual elements in game, concept arts)	(Visual) Aesthetics, sensual/sensory, presence (spatial), environment, immersion (sensory), and audio-visual
Emotional	Dramatic Elements (e.g. game characters, Narratives, Lore)	Stories/narratives, fantasy, expression, imaginative/imagination, fun, emotion, characters, and involvement
Socio-cultural	Game's social and cultural impacts (e.g. fan communities, reviews, advertising)	Society, community, contexts, frames, lenses, social interaction, presence (social), previous experiences, empathy, relatedness, and meaning

Table 3.1: Overview of the aspects of game-related experiences.

3.1.2 Explaining Types

ExpTypes reflect their most salient game "element"[3], making them intuitive and relatable, like:

Mechanical: Praising Celeste's [128] movement controls (i.e. mechanics) as being "tight";
Aesthetic: Feeling creeped out by the "vibes" of a horror game;
Emotional: Crying at the implications of putting our character's degree under the bed in Unpacking [528];
Socio-Cultural: Posting a TikTok compilation of our favourite character pairing.

However, we want to be clear about the meanings and implications of choosing Aesthetic, Emotional, and Socio-Cultural as type names since these terms can be overloaded. Here, **aesthetics** refers to the sensory experience derived from game audio-visual elements[4]. **Emotional** captures the empathetic and reflective relationship a person may have with the dramatic elements of the game at an individual level[5]. **Socio-cultural (SCX)** covers both the *social* elements of individual interactions (e.g. participating in online forums, stream chats, personal communications with other members of the game community) and the *cultural* elements of group actions and beliefs (e.g. gamer identities and politics).

[3]Game elements can be internal (e.g. mechanics, characters) or external (e.g. advertisements, fan communities) to the game itself.

[4]Aesthetics has been used in games to mean sensory phenomena, shared elements of an art form, and emotive-expressive acts [326]. Our focus on sensory phenomena sometimes overlaps with the meaning of aesthetics as the collective set of patterns for a particular art form but only because audio-visual elements and style are heavily coupled with this meaning (e.g. Cyberpunk aesthetic)

[5]In comparison to the ways mechanical and aesthetic experiences generate visceral emotions; a gut-response to stimuli. Emotional and socio-cultural experiences generate reflexive emotions that require context, knowledge, and reflection.

A Quick Aside...

We make a single category from social and cultural experiences as they are so interlinked in the literature we reviewed. However, we recognize this is an oversimplification of these experiences and foresee future iterations of this framework splitting these up.

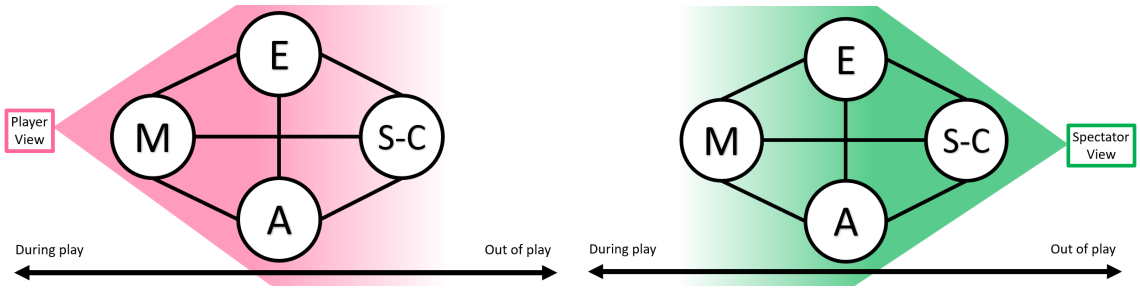
Connecting ExpTypes. Complex experiences arise from the interactions between ExpTypes (lines in Fig. 3.1). Consider the *mechanical-socio-cultural experience* of "deforestation" in Stardew Valley [93]. In the game, players need large amounts of wood, which they can get by chopping down trees. Players often run out of trees on their farm and turn to logging the nearby Cindersap Forest, which also quickly runs dry for wood-hungry players. The players are procedurally enacting deforestation. The game does not mechanically punish this behaviour as the forest's trees repopulate at the start of every season (except winter). There are also no social consequences as the villager's friendships points are unaffected by these actions (even the environmentalist, Leah!). The importance of wood almost implies that the player is actually *encouraged* to play this way. And yet, we see players on Reddit discuss the feelings of guilt they have over this logging and the ways they self-impose limits or offset this guilt through restorative forestry[6]. Some community members even suggested there should be an "anti-deforestation event" added to the game to impose mechanical consequences. As players understand and discuss the procedural rhetorics [49] of harvesting wood from the Cindersap Forest, their personal environmentalist values begin to affect how they understand and engage with the mechanics of the game. The player's SCX (discussing environmentalism in the game) changes their MX (how they interact with the mechanics).

Organizing ExpTypes. ExpTypes are naturally organized along a timescale from *during play* to *out of play*, indicating the dominant period for each experience. For example, MX (like building a hoverbike in Legend of Zelda: Tears of the Kingdom [346]) happen during play, and SCX (like posting a picture of your hoverbike on Reddit) happen out of play. The time-context of ExpType interactions are based on where they fall on the connecting line. Recall our example of the procedural rhetorics of Stardew's deforestation. This interaction is happening both during play and out-of-play, where the SCX influences future MX.

3.1.3 Lenses for ExperT

While we focus on PX, individuals can interact with games in multiple ways, including as spectators. ExperT captures the unique perspectives of games from these contexts through the *player* and *spectator* lenses. These lenses situate GX into a role-based context with clear timescales and characterizing experiences, thus making it easier to understand the differences between PX and spectator experiences (SX). We leave the player and spectator as black-box systems that bring their own personal contexts to the experience.

[6]Links to Reddit posts are not provided in order to preserve poster privacy.



(a) Player lens (b) Spectator lens

Figure 3.2: Analytical lenses for ExperT showing the ExpType which characterizes the lens experience and its dominant timescale. The colour strength proxies the abstraction of the lens experience based on distance from the characterizing experience.

Player lens (Fig. 3.2a) forms experience "during play", and is characterized by the game's MX. As the characterizing experience, MX is a barrier to engaging with other ExpTypes. Imagine a player who cannot master the mechanics of a Souls-like game and so quits at the first boss. Their inability to play, means they cannot engage with the complex emotional experiences of the game's environmental and emergent storytelling. This may even lead them to have a negative opinion of the whole game series, which they might voice online — a negative SCX.

Spectator lens (Fig. 3.2b) forms experience "out of play" and is characterized by the SCX due to the inherently parasocial and community building aspects of spectating [180]. While some SX are positive (e.g. the awe of watching a professional's skills), others can be negative (e.g. being bullied in Twitch chat). The SCX can be a barrier to further experiences. Consider a young woman, new to gaming, who wants to develop her fan knowledge of eSports. She starts to watch a top-rated streamer on Twitch, and musters up the courage to ask a question about the game, only to be met with sexist jokes about how she should be back in the kitchen. Feeling unwelcome she turns off the stream and will not engage again. This woman is barred from engaging with the deeper emotional and aesthetic experiences of watching this stream. This negative SCX characterises the rest of her interactions with the gaming community.

3.2 Uses for ExperT

We work through two examples, to show how ExperT is useful for analysing and orienting games-related work in ways that raise new questions and highlight gaps in research.

1. Comparing work on immersion: We visually compare models (Table 3.2) by placing them on ExperT to see which ExpTypes they cover and interact with (Fig. 3.3). This visualization highlights that *immersion* as a concept is heavily based on MX, and raises an obvious question: *are there socio-cultural elements of immersion*?

Paper	Immersion Components
Adams [7]	Tactical Immersion Strategic Immersion Narrative Immersion
Brown and Cairns [67]	Engagement Engrossment Total Immersion
Jennett et al. [212]	Cognitive involvement Real world dissociation Challenge Emotional Involvement Control
Ermi and Mäyrä [127]	Sensory-based Immersion Challenge-based Immersion Imaginative-based Immersion

Table 3.2: Comparing different immersion models by their components via ExperT

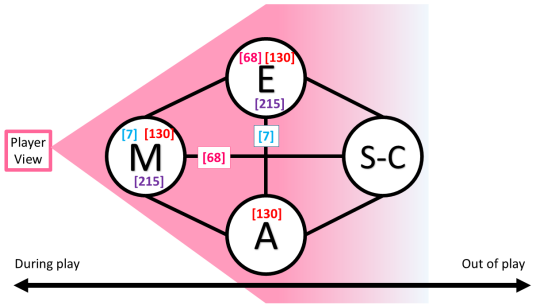


Figure 3.3: Colour-coded citations represent a component from the associated immersion model. Citations are placed on the Experiential Tetrad to show coverage of different experience types by immersion models.

2. Connected ExpTypes We place a sample of work about the interactions of ExpTypes (Table 3.3) on ExperT (Fig. 3.4). While it is not a comprehensive view of *all* games-related research, a quick glance shows an apparent mechanical bias in this sample. The empty lines of Fig. 3.4 beg to be filled by further research.

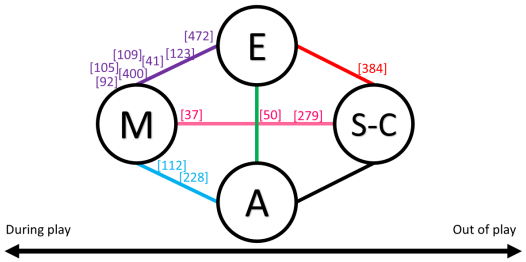


Figure 3.4: Placement of work on colour-coded experience lines, with reference numbers in matching colours.

ExpType	Example of Work
Mechanical-Socio-cultural	Accessible Play Experiences [36] Queer mechanics [276] Procedural rhetorics [49]
Mechanical-Aesthetic	Procedural game design [111] Mechanic/Aesthetic Genres [225]
Mechanical-Emotional	Affective gaming [396] Narrative mechanics [104] Emotional challenges [91, 108] Aesthetic gameplay patterns [40] Ludonarrative dissonance [121] Difficulty as aesthetic [468]
Emotional-Socio-cultural	Beyond empathy [381]

Table 3.3: Examples of work along different connections.

From two small examples, ExperT visualizes insights and raises several questions. We see MX are over-emphasized in individual and connection work. We can hypothesize this

is related to an implicit player-centric focus to the work, as the player-lens leans heavily on MX. This leads us to ask how would existing work look through a spectator lens? We see that work connecting ExpTypes focuses at most on two types. Finding work that looks at intersections of multiple aspects is difficult, with our effort resulting in Murphy's work on using ludonarrative dissonance for political commentary [314], and Thon's case studies of the aesthetics of horror [471]. From the sample we wonder whether this inherent two ExpTypes scope is because of the ongoing difficulty with describing, modeling, or measuring these interactions in a systematic way.

Quantifying experience. ExperT unifies PX and GX work into a clear framework. This sets us up for constructing specific measurable models for ExpTypes. As stated in Ch. 1, we scope our thesis to explore MX from the player lens. Given what we see from our examples, MX is more frequently studied and so we have more to draw from as we set out to explore it. We continue this chapter explicitly defining MX and setting up the theory for our work.

3.3 Mechanical Experience from a Player Lens

We propose:

Definition 3.1. ***Mechanical experience*** is the relationship between a ***player's ability profile*** and a ***challenge's competency profile***.

Where:

Definition 3.2. ***Player ability profiles*** describe the player's proficiency level for each of the possible human cognitive and motor abilities.

and:

Definition 3.3. ***Challenge competency profiles***[a] describe the ability proficiencies needed to complete the challenge.

[a] First mentioned in Ch. 1.2 and inspired by A. Fleishman, K. Quaintance, and A. Broedling [2].

With this understanding, we can directly measure MX by comparing player profiles to competency profiles (i.e. gameplay requirements) in a way that reflects the interactivity-as-demand paradigm from the Video Game Demand Scale [55].

Our measure-focused model defines two important qualities of MX: *mechanical achievability* (if the player can play the game) and *mechanical difficulty* (which game elements affect mechanical achievability). Understanding these two concepts would allow us to provide actionable design recommendations for specific gameplay. We expand on these concepts in the following sections.

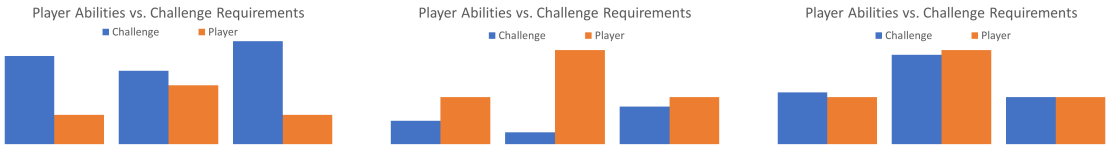
3.3.1 Mechanical Achievability

Consider Guitar Hero, a game series where players "play" songs by pressing fret buttons and strumming in time with the music and visuals. Just getting through the song (not getting a high score) requires motor dexterity in moving between the "notes" on the fret board, coordination to strum and move between the notes at the same time, and a sense of rhythm and timing to keep "on the beat" (see Fig. 3.5). For a player with arthritis or other motor issues, the gameplay may not be *mechanically achievable*.

Figure 3.5: Video clip of Guitar Hero gameplay from Youtuber GuitarHeroPhenom. Video shows player getting a 100% score on Expert mode. Click image to play.

Definition 3.4. ***Mechanical achievability*** describes the degree to which a player can succeed at a game based on comparing their ability levels (*ability profile*) to the challenge's requirements (*competency profile*).

Gameplay is easily achievable (maybe even boring!) when the player's abilities are greater than the challenge requirements (e.g. Fig. 3.6b). Gameplay is difficult to achieve (or impossible) when the player's abilities are significantly lower than the challenge requirements (e.g. Fig. 3.6a). This reflects Flow (e.g. 87, 103) with the idea of a Goldilocks-band ability-skill matching which creates achievable gameplay and thus positive MX (e.g. Fig. 3.6c).



(a) Difficult gameplay (b) Easy gameplay (c) Balanced gameplay

Figure 3.6: Mechanical achievability concept visualized based on the Guitar Hero example, with challenge requirements in blue and player abilities in orange. This is not real data.

With our MX model, we can quantify mechanical achievability as the difference between player abilities and challenge requirements. This allows us to find the soft error bars for achieavability, which describe the space where players with low abilities may succeed at a challenge through extra effort, practice, or luck.

Related work. Mechanical achievability is effectively a subset of playability (for definitions see: 130, 317, 406, 461) focused on game design through difficulty and mechanics, and so manifests in adaptive gaming and accessibility literature (e.g. 68, 543). We focus this quick review on *dynamic difficulty adjustment* (DDA) — techniques which change the game systems in response to player behaviour evaluated by some heuristic function/metric. As covered by Zohaib and Nakanishi [543], DDA work has explored:

- *probabilistic methods* for optimizing when to adjust game difficulty like probabilistic graphs (e.g. 533), statistical distributions/models like Gaussian distributions (e.g. 424) or Markov models (e.g. 70);
- *single or multi-layered perceptron neural networks* to map design parameters (e.g. placement of gaps to jump in a platformer) to emotional responses (e.g. feelings of challenge, frustration, or fun) and then generate game levels (e.g. 367, 430);
- *dynamic scripting* (e.g. 445), where enemy behaviours (i.e. attacks and strategies) exist as weighted rules in a script (where the weights determine how often the rule runs) and the weights are updated after every encounter with the player;
- *Game managers* which monitors and adjusts in-game resources based on player performance metrics and behaviours(e.g. the Hamlet System 199);
- *reinforcement learning* where adaptive AI is given a reward function to tune the gameplay based on desired outcome characteristics like even win-loss ratios (e.g. 464);
- *upper confidence bound for trees (UCT) and artificial neural networks (ANN)* where the ANN is trained by data from the UCT simulation, with the longer simulations creating more difficult opponents as the longer a UCT simulation runs the more "optimal" the opponents moves are — ergo the more likely the AI is to win (e.g. 265);
- *affective modeling using electroencephalography (EEG)*, where the system measures the player's emotional responses (e.g. excitement levels) through biometrics, and increases difficulty or new game events when prolonged boredom is detected (e.g. 297, 449); and
- *self-organizing systems wih ANN* which treat game entities like NPCs as independent agents (hence self organizing) that each use an ANN to create their unique behaviours and strategies (e.g. 123, 356).

Commercial examples of DDA seem to implement game managers (e.g. Difficulty Scale in Resident Evil 4 [76], The Director in Left 4 Dead [493], Hamlet System in Half-Life 2 [491]), and simple dynamic scripting (e.g. rubber banding in Mario Kart [328]). Overall, DDA focuses on adjusting perceived or assumed sources of difficulty based on designer knowledge. Existing techniques do not address *why* players are failing, and is not clear whether players then receive the same designed MX. DDA work is also plagued by community views that adapted experiences are less valid (e.g. discussions in online forums like r/truegaming [500, 502], and in specific game forums if DDA is suspected [499, 503]). Our work aims to understand why players are failing, quantify their difficulty and accounted for it at a design level.

3.3.2 Mechanical Difficulty

Consider Guitar Hero again; difficulty levels are set by adjusting how many notes need to be played, and how quickly they move across the screen. The implicit dexterity requirements correlate to how fast the icons move along the virtual fretboard and the number of buttons needed to be pressed. This illustrates how designers craft a challenge's mechanical achievability through choices about the structure and tuning of mechanics and game elements.

Definition 3.5. *Mechanical difficulty* is the relationship between *challenge design decisions through gameplay elements* and the *challenge competency profile*; it shows how changes in the former express themselves in the latter.

With our MX model, we can highlight the abilities that bottleneck player performance, and trace them to the specific design decisions. This mapping gives us insight into *why* players fail, and indicates *what* accommodations a player needs to succeed. It also allows us to consider how tweaks to the bottlenecking element changes the overall shape of the competency profile. This is useful when you have multiple abilities that are close ties for the bottleneck; adjusting the bottleneck alone might create new sources of difficulty in other areas.

Related work. As far as we know mechanical difficulty has not been specifically studied. There is a lot of larger interface knowledge from HCI that provides similar guidelines (e.g. how changing target sizes improves accuracy by reducing stress on player's abilities by Fitt's Law), but it ends at the interface and interaction elements. We take an adhoc approach to mechanical difficulty, attempting to identify it on a challenge by challenge basis, with the intent of validating our sources of mechanical difficulty experimentally.

3.4 Conclusion

Having constructed ExperT we scoped our work to focus on MX through a player-lens. We further refine and reframe our understanding of a player's MX as the relationship between *mechanical difficulty* and *mechanical achievability*. Embedded in this reframe is the idea of a measurable model of MX based on *player ability profiles* and *challenge competency profiles*. Moving forward, we need to further flesh out the requirements and design decisions for this ability-based MX model.

Take home points

From this chapter we learned the following meta-lessons:

- ExperT is useful for analysing and orienting existing work to highlight research gaps.
- ExperT gives us clear scopes for developing quantifiable models of experience.
- MX (in the player-lens) is the relationship between player abilities and challenge requirements, quantifiably modeled by mechanical achievability and mechanical difficulty.

Chapter 4

Modeling Mechanical Experience

Having defined mechanical experience (MX) and its components in Ch. 3, we start to tackle modeling it. Since MX is the relationship between a player's abilities (i.e. player profile) and challenge requirements (i.e. competency profile), an MX model requires a player model and challenge model. These three models are highly coupled, but will be somewhat independently designed and presented. So it is important that the guiding requirements and design decisions for this large project are clear.

For this chapter, we immediately jump into presenting the requirements for our model. We derive these requirements from examples and case studies of game experiences our final model should differentiate. We then specify design decisions we make about the structure of our models. Finally, we present the design of *jutsus*: our knowledge capture artifact that combines the player, challenge, and MX model.

4.1 Requirements of Our Models

While broadly we look to understand MX, mechanical achievability (MechA) and difficulty (MechD) we specifically care about how cognitive and motor abilities are used to interact with various challenges (RQ1). So our models must let us easily discuss the player, challenges, and their relationship in terms of cognitive and motor abilities (formalised as R1).

Requirement 1. Models of challenge and player must allow for discussion in terms of cognitive and motor abilities.

RQ1 specifically targets the interaction of these abilities with the challenges. Interaction here means how these abilities are used to complete the challenge with varying degrees of success. A model of MechA (Def. 3.4) can answer RQ1 since it relates the player model to the challenge model. For simplicity, these three models must be be created using the same underlying terms (formalised as R2).

Requirement 2. The model for mechanical achievability must be in the same terms as the challenge and player models to make their relationship easier to analyse.

We focus on requirements for the player and challenge models since they are the base of MechA. We formalise the basic goal of our models in R3.

Requirement 3. All models should be able to differentiate between meaningfully different instances.

For players, this means there should be a notable difference between players who have different skill levels. This can be handled in a straightforward way through skill level measurement. But how would we differentiate challenge instances (i.e. how would this model show Challenge 1 to be different from Challenge 2?). Our natural instinct is to consider instances as "different" when they "feel" distinct. However what does it mean to "feel" different in the context of a video game challenge? Consider the following cases:

Case 1. We have two games, Portal [492] and Halo 3 [69]. Both games require moving through a set level in a first-person perspective using a gun to interact with the environment. They both heavily focus on aim-and-shoot as the main gameplay mechanic. However Portal is a puzzle game where you are aiming your gun to place portals to solve puzzles and traverse through the level, while Halo 3 is a combat focused game where you are aiming at moving enemies to clear the field and defend locations. These games use similar mechanics, mechanisms of interaction, and camera contexts but provide very different gameplay experiences because they are fundamentally asking you to do different things. Portal is slower and requires more cognitive abilities related to spatial understanding, while Halo 3 requires faster reflexes and more motor control/focus.

(a) Video clip of Portal Test Chamber 13 by Youtuber p2wiki. Click to play.

(b) Video clip of Halo 3's 3rd mission (Crow's Nest) by Youtuber Vezuvius. Click to play.

Figure 4.1: Comparison of Portal and Halo 3 gameplay to illustrate Case 1. Image shows the aesthetic similarity between games, while clicking on the video shows the significant differences in gameplay as described in the case study.

Case 2. Consider the Protect the Milk side quest from Legend of Zelda: Majora's Mask [331], and Pokémon Snap! [176]. Again, these are two games that use aim-and-shoot mechanics from a first-person perspective. Both games automatically control the player's movement (i.e. on-rails), so the player only has to consider looking at the moving targets. However, Majora's Mask is combat focused as you shoot arrows at your pursuers, while Pokémon Snap!

has you taking pictures of Pokémon. Both games require strong perception and tracking skills, and rely predominantly on reaction time. Neither requires more or different skills than the other.

(a) Video clip of the Protect the Milk side quest in Majora's Mask by Youtuber OddlerPro. Click to play.

(b) Video clip of Pokemon Snap's Beach course from Youtuber JackLiberty0 Gaming. Click to play.

Figure 4.2: Comparison of Majora's Mask and Pokemon Snap gameplay to illustrate Case 2. Image shows difference in appearance between games, while clicking on the video shows the mechanical similarities in gameplay as described in the case study.

In Case 1 (Portal/Halo 3), we have two games that look similar but the underlying challenge is different. In Case 2 (Majora's Mask/Snap), we have two games that look different but the underlying challenge is the same. Our challenge model should be able to identify cases like these to determine when challenges are the same or different based on their ability requirements. We formalise this in R4.

Requirement 4. The challenge model should differentiate challenges based on the competency profile regardless of the aesthetics.

But what about challenges that use the same abilities in different amounts? Consider the following cases:

Case 3. Consider the roller coaster balloon segment in Super Mario Sunshine [329], and the New Pokemon Snap! [28]. Both games are on-rails (the movement controlled by the game) and require aiming and shooting at targets (shooting balloons in Sunshine, and taking pictures of Pokemon). The games are different in their presentation (Mario is third person, Pokemon is first person) and in speed (Mario's roller coaster moves faster than the Neo-One ship in Pokemon). Here the difference between these challenges are specific design context variables: the speed of motion, and position of the camera. These differences mean the proficiency in the underlying skills needs to be higher for Mario as the speed means there is less time to react.

(a) Video clip of roller coaster segment in Super Mario Sunshine by Youtuber hyrulesmasher. Click to play.

(b) Video clip of New Pokemon Snap's Beach course from Youtuber FCPlaythroughs. Click to play.

Figure 4.3: Comparison of Super Mario Sunshine and New Pokemon Snap gameplay to illustrate Case 3.

While we can explain *why* the MX in these cases are different, we do not have a good universal call on *when* they become different to a player. This is a question we may not be able to incorporate as a design decision or requirement since it could be unique to each challenge type. However, we know the relationship between gameplay elements and ability requirements from MechD (Def. 3.5), and could use this to we decide where the line is between "same" and "different" for specific challenges. We make including MechD information a requirement (R5) so that we have a starting point to empirically differentiate challenge instances later.

Requirement 5. The challenge model must include mechanical difficulty information to help differentiate between challenges.

4.2 Design Decisions for Our Models

Falling out from R1 and R2 is the understanding that MechA must also allow for the discussion of cognitive and motor abilities. This means our three models (player, challenge, and MechA) should be written in terms of human cognitive and motor abilities (our first design decision, DD1). Therefore we will need to explore and curate a list of human cognitive and motor ablities from research in other areas.

Design Decision 1. The player and challenge models will be built from a single set of cognitive and motor abilities.

Considering R3, we need to define the scope and scale of the abilities in this set. Imagine a person interacting with a computer via keyboard and mouse. While typing and clicking are different activities with different uses, underlying them is the ability to press a button. So

if our ability set is scoped to the level of activity we may have a lot of unnecessary overlap, especially when you consider how many activities we could find (just look at all the verbs in English!). This leads to another design decision formalized as DD 2.

Design Decision 2. Our models should be constructed from a set of basic abilities that can be used in various combinations to perform different activities.

Combining R3 and R4, we must also consider how many abilities we may find that would have no current use in a video game context. Current controllers do not capture movements like toe wiggling or breathing as inputs, indicating they are not effective controls for gameplay. However, the scope of abilities used in video games will change as input technology changes (you can now play Flappy Bird with your kegels using a PeriFit). For now, we choose to focus on current gaming contexts but want to leave this open to expansion.

Design Decision 3. The ability set should be scoped to current game playing contexts in a way that allows other abilities to be added later.

R4 and R5 say the challenge model needs to differentiate between challenges via competency profile and MechD. Since our models are effectively subsets of the cognitive and motor ability set, we can leverage measurable differences in the ability proficiency levels to address these requirements:

Design Decision 4. Challenges will be considered distinct when they have a different set of abilities or observable differences in their ability requirements.

To use ability requirements/proficiency to compare challenges (R4) we need to think about how they will be represented. Consider the gameplay of Portal, which requires precise controller inputs (motor abilities) and complex problem-solving (cognitive abilities). The measurement methods, units, and scales of these abilities are inherently different and so difficult to compare. As well, the precise controls and complex problem-solving are actually a concurrently working set of even more basic abilities whose performance combines into what we see on screen. It would be ideal of ability measurements to be normalized in a way so direct comparison between abilities can be made, and any combinations of multiple ability tests could be combined without competing units.

Design Decision 5. Ability measurements in the models will be represented through a normalized unit to allow for multiple measurement methods and comparison between ability proficiencies in the same model instance.

4.3 Designing Jutsus

In outlining the meta-requirements and design decisions for our work, we start to visualize what these models will look like and how to present them. We now put our three models

(MX, Player, and Challenge) together into one designed knowledge capture artifact: jutsus.

Definition 4.1. Jutsus are a structured representation of a specific *mechanical experience* made from combining a *challenge model* and a *player profile*.

Jutsus have three sections: challenge description, player profile, and mechanical experience analysis. Each captures an instance of the associated model. We outline our design for each section below, as a preview of the models and what jutsu can be. We fully explore jutsu in Part IV.

A Quick Aside...

In this thesis the terms model and profile are used to differentiate between the abstract idea and a specific instance. For example, the Player Model refers to the larger idea of the ability set which defines a player, while a Player Profile describes an individual instance of a player model (i.e. a measured representation of a player's abilities). Keep this in mind as these terms may be used at different points to reference specific things.

4.3.1 Challenge Description (Challenge Model Instance)

An individual challenge is described in four parts: the definition, the mechanics, the context, and the competency profile. Fig. 4.4 is an example of a challenge description for Single Input Button Mashing challenges. Details about challenge construction are found in Ch. 14.

Challenge definition is a one sentence, natural language description of the challenge's goal and key mechanics. The purpose of this is to easily summarize the challenge, abstracted from any individual game example or aesthetic. It also highlights key terms that are important to understanding the challenge.

Mechanics are then described as succinctly as possible. Mechanics are the formal rules of the challenge, and so relate specific inputs to different in game factors. Alongside the mechanics we present the variable components in the game design which influence the mechanical difficulty.

Context describes the other factors that influence a challenge's play and feel. We specify four elements: the mechanism of interaction, controller type, number of players, and type of play. The *(mechanism of) interaction* describes the motor action that needs to be done to complete the challenge; this is dependent on the input device. The *controller type* describes the specific input device used. The same mechanism of interaction may exist across various controller types. The *number of players* captures whether the game is single or multiplayer. The *type of play* describes whether the game is co-operative, competitive, or mixed; it also separates these options into whether it is team-based or solo-based for these.

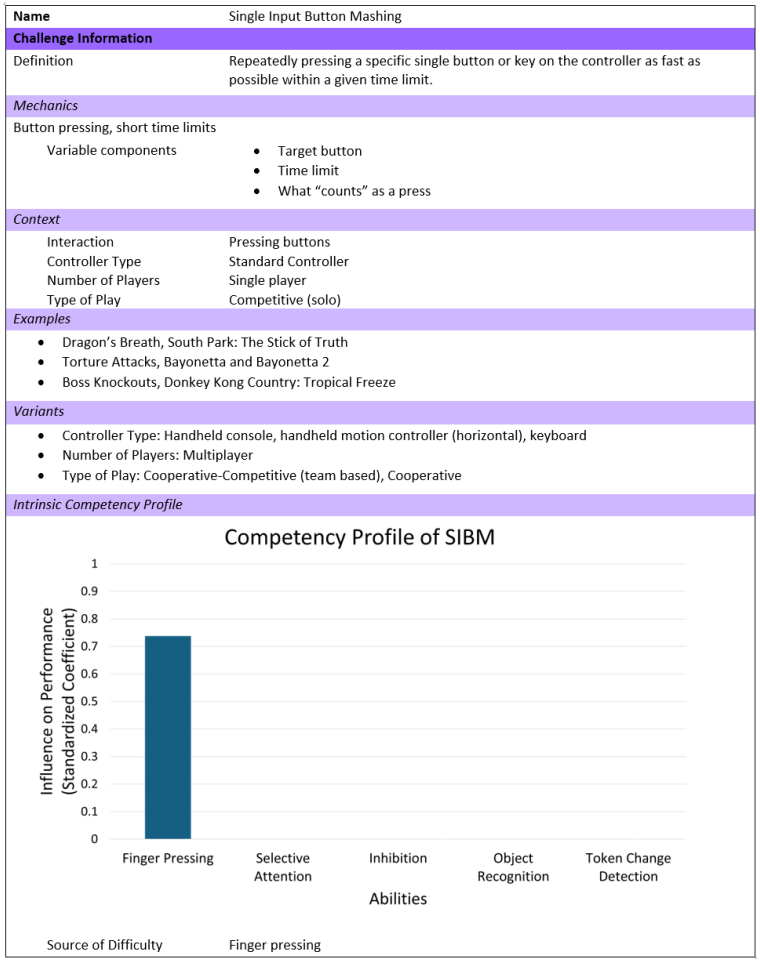


Figure 4.4: Challenge description segment of Single Input Button Mashing Jutsu. Covers the challenge definition, mechanics, context, and competency profile.

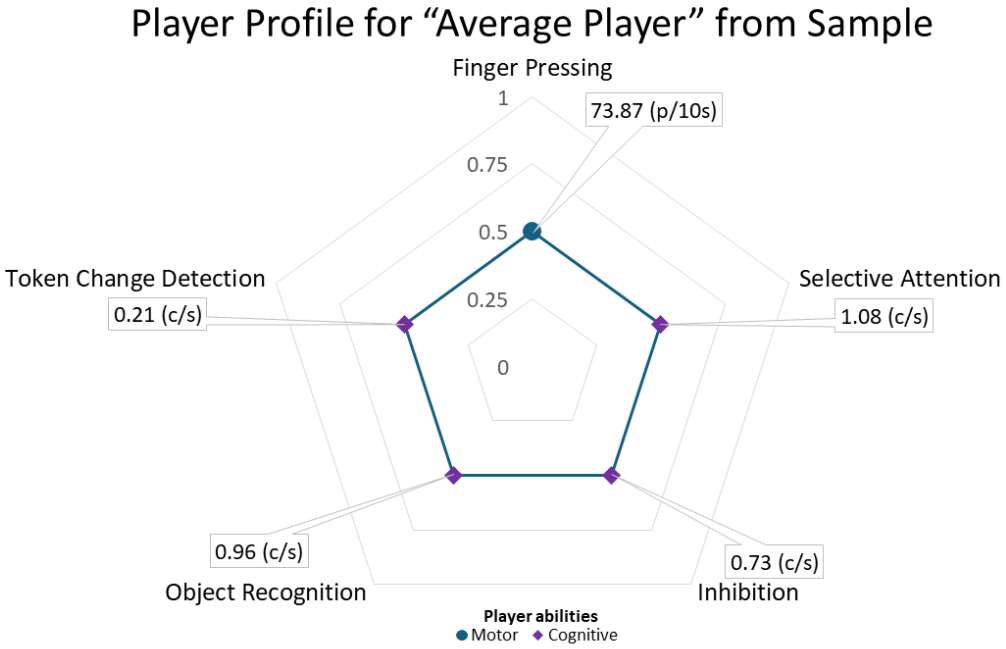
Intrinsic Competency Profile is the graphical representation of cognitive and motor abilities needed to successfully complete the challenge. Each ability is represented by a bar in the graph and is assigned a normalized value between 0 and 100 (scaled down to range 0-1). The value of the ability represents proficiency needed in that ability to complete this challenge.

As we have not yet designed the abilities or tests to measure them, we believe the relative proficiency levels are more important than the specific value of an ability. 0 means an ability is not used at all; 1-25 means that the ability is used, but not important; 26-50 means the ability is noticeably used; 51-80 means the ability is the limiting factor of play. The reason that specific values are given instead of just the range is to understand the relative importance of abilities that fall into the same range.

4.3.2 Player Profile (Player Model Instance)

A graphical representation of the player's ability levels, where proficiency in each ability is measured independently, and then normalized on a scale of 0 to 1 for easy visualization

and comparison. In our current work, these profiles represent the player in relation to their sample population. Ideally, with larger sample sizes and normative data these profiles begin to approximate a player relative to the general population. Fig. 4.5 is an example of an "average" player from our empirical studies in Ch. 16. Details about how we construct profiles can be found in Ch. 12.1.



Legend p: Button presses, c: Correct Responses, T: total number of trials, s: Seconds

Figure 4.5: Example Player Profile of "Average Player" from Competency Studies (Ch. 16). Blue dots indicate motor abilities, purple dots indicate cognitive abilities.

4.3.3 Mechanical Experience Analysis (MX Model Instance)

The analysis section of the jutsu structure covers the mechanical achievability graph, potential sources of difficulty, and suggested tweaks.

Analysis graph compares the challenge's specific competency profile and player profile (Fig. 4.6). Details about how we construct graphs are in Ch. 20. Based on the challenge details, the graph calculates the minimum ability levels needed to reach the goal (represented by gray dotted line) and the potential rank[1] associated with each proficiency level. The player abilities are then plotted to show whether they meet the requirements (MechA), and what rank they would approximately receive. This visually indicates whether players will struggle with the game (i.e. poor mechanical achievability).

[1]Score buckets relative to the goal

Potential sources of difficulty are then outlined based on the particular abilities that are likely to be a problem for the player.

Suggested tweaks are then presented based on the variable components that address the areas of difficulty.

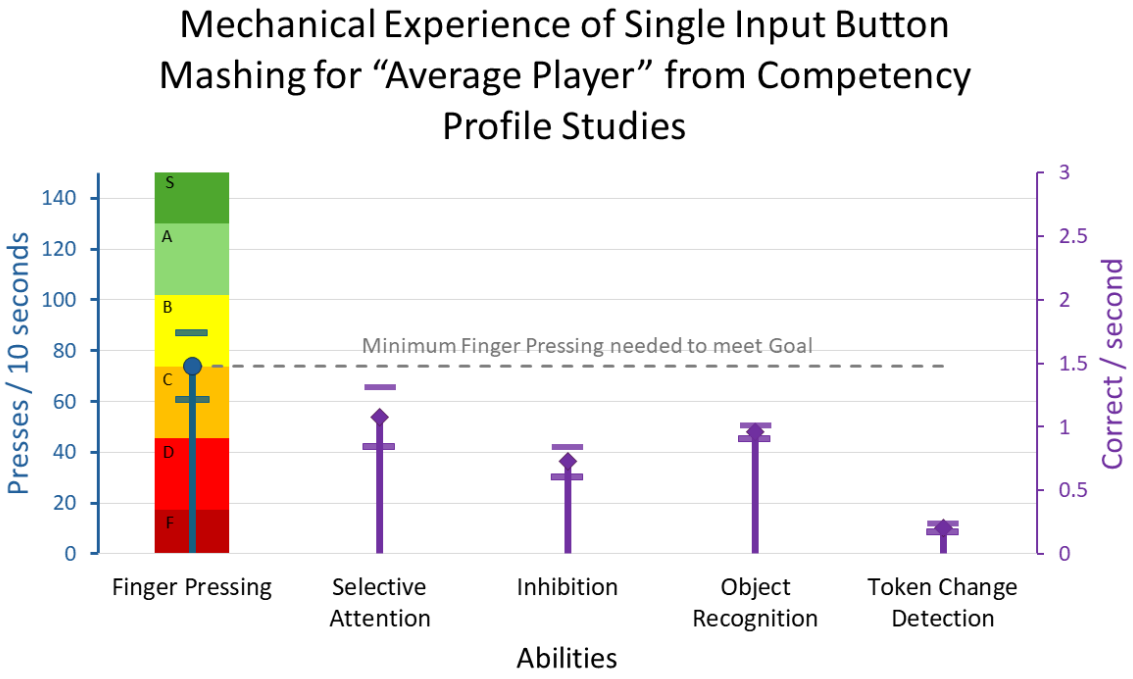


Figure 4.6: Single Input Button Mashing Analysis graph for Average Player from Competency Profile Studies.

4.4 Wrapping up on Modeling Mechanical Experience

We outline four major requirements:

R1 the challenge and player models should allow for discussion of cognitive and motor abilities;

R2 the models for mechanical achievability, the player and challenges must be constructed in the same terms to make their relationship easier to analyse;

R3 the models should differentiate between meaningfully different instances;

R4 the challenge model needs to distinguish between challenges that are the same or different based on their competency profiles; and

R5 the challenge model must include mechanical difficulty information to make the differentiation easier.

These inspire five design decisions:

DD1 the models for the challenge and player will be made from the same set of cognitive and motor abilities;

DD2 the cognitive and motor abilities will be as atomic as possible so we have a minimal set that can be used to describe how to complete any action;

DD3 the cognitive and motor abilities set will be scoped to those that are useful in a video game playing context;

DD4 challenges will be considered distinct when they have a different set of abilities or observable differences in their ability requirements; and,

DD5 proficiency levels for each of ability will be presented as a normalized unit.

We summarize their relationship in a traceability matrix (Table 4.1).

	DD1	DD2	DD3	DD4	DD5
R1	✔				
R2	✔				
R3		✔	✔		
R4			✔	✔	
R5					✔

Table 4.1: Traceability matrix for requirements and design decisions.

The requirements and design decisions shape the direction of our work, and lead to our design for jutsus. Jutsus are specific models of MX for a given challenge and player. Jutus represent a quantitative model of MX and so can address our larger research questions and thesis goals. While we foresee many jutsus existing, our current focus is to flesh out and prove the idea through establishing the models (Part II and III) and constructing a set of example jutsus (Part IV).

Take home points

From this chapter we learned the following meta-lessons:

- We are constructing three models: a Player model, a Challenge model, and a mechanical experience model.
- Our models are built from a single set of human cognitive and motor abilities contextualized to gaming.
- Our models combine into a jutsu which can tell us a particular MX for a given challenge and player.

Chapter 5

Closing Remarks: The Setup

We close out Part I with a better understanding of the goals of our thesis, and an approach for answering our research questions. We explored contemporary understandings of player experience through a narrative review (Ch. 2) which led to the creation of the Experiential Tetrad (Ch. 3). We use the Experiential Tetrad to scope our work to focusing on the mechanical experience of a game through a player-lens. We construct mechanical experience as the combination of mechanical achievability and mechanical difficulty, which we can directly tie to RQ1 and RQ2. We build on this understanding by creating requirements for modeling mechanical experience and proposing a presentation of that information through jutsus (Ch. 4).

What we learned in this part:

- Game-related experiences are complex multi-dimensional phenomena, and so are hard to define and model precisely.
- We can use the Experiential Tetrad to explore existing work on game-related experiences and create novel work.
- The Experiential Tetrad allows us to scope our thesis to studying mechanical experience (experience from interacting with the gameplay) through a player lens.
- Mechanical experience can be modeled as the relationship between a player's abilities and challenge requirements using mechanical achievability and mechanical difficulty.
- To quantifiably study mechanical experience and its components, we need a player model and challenge model based on human cognitive and motor abilities.

What we produced in this part:

- The Experiential Tetrad (Fig. 3.1) as a way to view different aspects of game-related experiences;
- Working definitions of mechanical experience (3.1), mechanical achievability (3.4), and mechanical difficulty (3.5);
- The definitions of player profiles (3.2) and challenge competency profiles (3.3);
- Requirements (4.4) and design decisions (4.4) for modeling mechanical experience; and,
- A high-level concept of Jutsus that brings the pieces together (Ch. 4.3).

Part II
The Player Model

Here we tackle creating a player model that can be used in assessing the mechanical experience of a gameplay challenge. To do so we explore existing player and user models in games research and human-computer interaction to see if they fit our needs (Ch. 6). We decide to create our own player model, and so spend the next chapters compiling lists of relevant motor (Ch. 7) and cognitive (Ch. 8) model. Having determined the abilities in our model, we need a way to measure them so we can construct player profiles (Def. 3.2). We explore how our model abilities are measured (Ch. 9), and use this information to create a battery of ability tests in the form of mini-games (Ch. 10). We attempt to show convergent validity of our mini-games to existing measurement methods through a correlational study (Ch. 11), though the results are mixed. We end this part by synthesizing everything we have learned in this part to construct player profiles (Ch. 12).

Chapter 6

Player Profiling in Theory

Recall from Ch. 4, we are constructing a Player model based on cognitive and motor abilities available in a gaming context. Before we start compiling abilities for our model, it is prudent to explore existing work on player modeling. This way, we can see if there are existing models that could serve our purpose. Ideally, the model we are looking for should take an *ability-based design* approach, such that users are represented as an *ability model* which defines context relevant abilities. However, Nolte et al. [349] note as of 2022 that this area is largely neglected especially for a generalizable model.

Player modeling generally refers to work on player typologies, which segment potential players into types based on their motivations and behaviours (e.g. Bartle [30] and Yee [534]). It could also reference player trait theories, which model players along a set of standard traits (e.g. 318). There are many analyses and criticisms about the usefulness and validity of player types (e.g. 34, 178). For us, these models do not seem appropriate because of their level of abstraction (predicting *what* a player will like/want, not specifically *how* they will accomplish the gameplay). Despite their unlikelines to produce a ready-to-use model, we broadly review user and player models to get a sense of how our work fits into the larger domain.

This chapter presents a narrative review of models that may be applicable to interacting with a video game. We first review basic user models from human-computer interaction (HCI). These represent a historical look at user modeling; they are much more performance-focused and computationally designed than player-specific models. We then turn to player typology work from games user research (GUR) and games studies (GS). We separate these models into psychographic models and behavioural models as per two meta-reviews: Bateman, Lowenhaupt, Nacke, et al. [34] and Hamari and Tuunanen [178]. However, this distinction is somewhat arbitrary as models frequently overlap or are ported between categories. We provide more details for these models due to their player focus, and to highlight the overlaps between models. We end with a synthesized list of takehomes from this literature review.

A Quick Aside...

Two personality models frequently feature in this area: the Myers-Briggs Type Indicator model (MBTI) [62] and the Five Factor Model (FFM, Big 5, or OCEAN) [97]. MBTI assesses people along four spectrums (**E**xtroversion-**I**ntroversion, **S**ensing-i**N**tuiting, **T**hinking-**F**eeling, and **J**udging-**P**erceiving) creating sixteen possible types (e.g. INTJ, ESFP)[a]. The Big 5 assesses people along **O**penness to experience, **C**onscientiousness, **E**xtraversion, **A**greeableness, and **N**euroticism[b]. MBTI is generally criticised for poor scientific rigour, stability, and validity [450]. The Big 5 has more empirical evidence for validity, stability, and rigor [288].

[a] Find your Myers-Briggs at 16personalities.com.
[b] Find your Big 5 traits at bigfive-test.com.

6.1 Computational User Models (HCI)

HCI user modeling focuses on predictive user behaviour and performance modeling. Broadly HCI provides *personas* and *user case studies* as practical tools to form specific user models for individual applications. However, we are more interested in the generalizable user models from HCI research. These apply cognitive modeling for the purpose of task analysis. We cover three categories: GOMS models, performance models, and cognitive architectures.

Goals, Operators, Methods, and Selection rules (GOMS). Card, Newell, and Moran [80] present the *Model Human Processor* (MHP) as an underlying generic cognitive structure, which divides a person into three independently processing subsystems: motor, cognitive, and memory. This view of cognition supports constructing user models as a set of *goals* (things to be achieved), *operators* (context-appropriate actions), *methods* (sequences of goals and operators used to complete tasks), and *selection rules* (how the user chooses the right methods) [78, 80]. This GOMS approach has been widely used in HCI work, including for modeling games (e.g. 215), and is implemented in multiple ways. Two prominent GOMS models are the Keystroke Level Model (KLM) [79] and CPM-GOMS (cognitive, perceptual, and motor operators in a critical-path method) [213, 214]. GOMS limitations, as discussed by many (e.g. 78, 357), include its:

- focus on skilled users and errorless performance,
- lack of clarity on specific cognitive processes,
- assumption of serialized tasks (though this is addressed by CPM-GOMS), and
- inability to account for individual differences, fatigue, and mental workload.

GOMS conceptually aligns with our goals (modeling ability requirements of specific game tasks), but as it stands cannot fulfill them due to its scope-based limitations.

Predictive Performance Models. A significant body of work focuses on mathematically modeling human performance based on information processing theory (see Wickens et al. [518] or Ch. 8.1 for details). Card, Newell, and Moran [80] introduced two such models

to HCI: Fitts' Law [139] (for motor performance), and the Hick-Hyman Law [187, 201] (for choice-reaction time). While Fitts' Law has thrived in HCI work (e.g. 5, 270, 320, 321), Hick-Hyman has not [427]. This could be due to Fitts Law being standardized in the ISO-9241-411 procedure[1] for evaluating non-keyboard input devices [125], which has led to significant work evaluating different interfaces that continues to this day (e.g. 399). While these models help quantify general task performances in terms of execution times, they do not give us a holistic model of the user.

Cognitive Architectures (CA). These are generic models of cognitive structures and processes that can predict human behaviour and performance at any task [456]. CA are either *symbolic* (top-down process using high-level symbols for information processing), *emergent*[2] (bottom up process using low-level signals), or *hybrid* (combining symbolic and emergent) [119, 160]. Games-related HCI explores CA for believable agent behaviour (e.g. 71, 89, 246, 247, 248, 494), and user modeling to test specific cognitive abilities (e.g. 154, 260). Popular models in games-related HCI are the symbolic State, Operator and Result (SOAR) [249] (used in 246, 247, 248, 494), and the hybrid Adaptive Control of Thought - Rational (ACT-R) [13] (used in 154, 260). We also explore Executive Process-Interactive Control (EPIC) [229–231], Learning Intelligent-Distribution Agent (LIDA) [143, 439], and Connectionist Learning with Adaptive Rule Induction On-line (CLARION) [455, 456] to get a general sense of CA. CA generally focus on memory and attentional control, with less clear models of perception and motor control. Any CA would need significant extension to model the player in the way we want for the purposes we want, making them less appealing than custom creating our model.

6.2 Behavioural Player Models

In-game behaviours act as proxies for player gaming preferences and motivations. Ergo, behavioural typologies reflect different player needs, wants, and motivations. These models provide a useful vocabulary for designing and balancing gameplay for different player desires and experiences. However, they are not particularly useful for modeling how players perform individual gameplay at an ability-level. Some quantitatively model specific gameplay behaviours (e.g. 117), but are then highly tuned to a single game. We focus on common typologies, including: Hardcore-Casual, Robin Laws Gamer Types, Park Associates Gamer Types, Bartle Types, Yee's Motivations/Gamer Motivation Profile, and Drachen's player types.

Hardcore and Casual are opposing gaming behaviours, demographics, and philosophies, which represents both a play style and identity [286, p.40][3]. Historical work presents the hardcore gamer as a young, single, man who spends the majority of his freetime playing specific types of games[4] (e.g. 145, 223), despite ample contradictory academic (e.g. 170, 243,

[1] Previously ISO 9241-9
[2] Also known as connectionist
[3] We discuss gamer identities and its relationship to hardcore in Ch. 18.1.
[4] MMORPGs, FPS, RTS, and Sports games.

509, 523) and census-based (e.g. 15, 16) evidence. Casual represents everything hardcore is not; they approach games in a casual manner [244], play for enjoyment and are less tolerant of difficulty and frustration [7, p. 75]. Juul [219, p.54] notes the lack of evidence for this stereotype, and proposes instead that hardcore and casual are differentiated by their flexibility in playstyle and preferences: casual players have inflexible time demands and so require flexible gaming, while hardcore players are flexible in their time demands and so can play inflexible games. Researchers and players alike generally agree that the Hardcore-Casual dichotomy lacks clarity in meaning, and usefulness in exploring player behaviours. Recently, Brett and Soraine [60] examine the current behavioural practices of "Hardcore" and "Casual" and propose an extended typology of overlapping behaviours.

Laws [256] presents seven archetypes for tabletop role-playing game players. The *power gamer* (i.e. min-maxer) aims to optimize their play through strict adherence to and manipulation of the rules. The *butt-kicker* just wants to engage in combat without significant difficulty. The *tactician* focuses on strategic play within the rules. The *specialist* only cares about their favourite playstyle and does not care about the rules, systems, or world outside of it. The *method actor* cares about expression through their character. The *storyteller* cares about the narrative and will get bored if it moves too slowly. The *casual gamer* cares about socializing with the group, and do not have preferences about playstyles. Laws notes that individuals may present traits of more than one type. Work on interactive digital storytelling uses this typology to consider plot hooks for generative stories (e.g. 369, 472). These types are anecdotally derived, and unvalidated.

Bartle Types [30] outlines four personalities found in a multi-user dungeon (MUD): Achievers, Explorers, Socialisers, and Killers. These types describe players along two spectrums: acting vs. interacting, and players vs. the world. Achievers want to *act on the world* by gathering achievements and accolades which display their mastery at the game. Explorers want to *interact with the world* by discovering secrets and building a deep understanding of the game. Socialisers want to *interact with players* by building communities. Killers want to *act on players* by exerting control over others to induce negative experiences. This work is explicitly for designers to show how these player types are balanced in a healthy online ecosystem, and was not intended to be applied outside the context of MUDs and their descendants (i.e. MMOs) [31].

Cai [73] and Park Associates [361] present six gamer types as a market segmentation tool. *Power gamers* fit the "hardcore" stereotype. *Social gamers* use games as a means of connecting with others. *Leisure gamers* are hobbyists who play a lot of casual games, but prefer challenge and show more interest in new gaming services. *Dormant gamers* are enthusiasts who do not have time to game; they enjoy socialising and prefer challenge. *Incidental gamers* play to alleviate boredom, and play frequently but do not have larger defining traits. *Occasional gamers* focus on puzzle, word and board games. Power, leisure, and dormant gamers align with "hardcore" while social, incidental, and occasional gamers seem more "casual". While less popular than others here, this typology sees occasional mention in literature (e.g. 235, 526). These types focus more on *gaming* behaviour, rather than in-game

desires/behaviours.

Drachen, Canossa, and Yannakakis [117] extract player models from Tomb Raider: Underworld [102] by clustering game metrics using an unsupervised learning model of self-organizing maps. This results in four player types: *Veterans* (low total death count, deaths caused by environment, fast completion time, low to average help requests), *Solvers* (frequent deaths, deaths caused by falling, long completion time, help requests are minimal to none), *Pacifists* (variable total deaths, deaths caused by opponents, below average completion time, minimal help requests), and *Runners* (frequent deaths, death by opponents and environment, fast completion time, varied help requests). These types show different play styles throughout the game, and serve to show whether players are engaging with the game as intended (i.e. expected behaviour). While quantitative and useful, it is hyper-specific to Tomb Raider and developing further models would require significant expert knowledge of each game being analysed and the ability to pull meaningful game metrics.

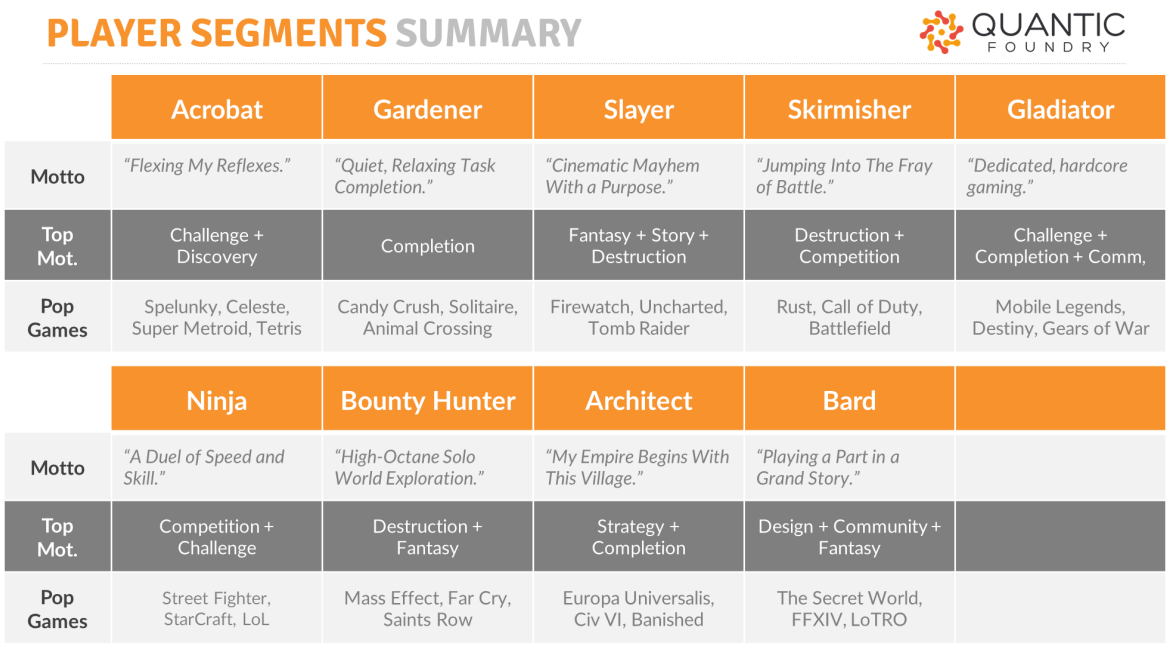
Gamer Motivation Profile [535] develops Yee Motivations [534, 536][5] further into a six motivation model, where each motivation is supported by two specific traits. The motivations are: action (*excitement, destruction*), social (*collaboration, competition*), mastery (*strategy, challenge*), achievement (*power, completion*), creativity (*design, discovery*), and immersion (*story, fantasy*). These motivations support nine archetypes (Acrobat, Gardener, Slayer, Skirmisher, Gladiator, Ninja, Bounty Hunter, Architect, and Bard) whose descriptions, primary motivations, and example games are summarized in Fig. 6.1.

6.3 Psychographic Player Models

Psychographic types segregate players by their "attitudes, interests, values, and lifestyles" [178], often using personality as a proxy. Like behavioural models, these are meant to capture player wants, needs, and motivations. Psychographic models are generally too abstract for our purposes. We focus on a sample of popular typologies: DGD1/DGD2, BrainHex, Unified Model, and Hexad.

Demographic Game Design Model (DGD1) [33] presents four player types and their matching MBTIs. *Conquerors* (TJ) are goal-oriented players who care about winning quickly but do not care about stories. *Managers* (TP) are mastery-oriented players who enjoy strategic and process-based challenges so long as they are not too difficult. *Wanderers* (FP) are novelty-seeking players who are emotionally-invested in characters but will give up on games they find boring. *Participants* (FJ) are social players who seek emotional connection with stories and characters, and will persevere through boring or difficult gameplay to get this. DGD1 has been tested against the Big 5 traits [294]. Bateman, Lowenhaupt, Nacke, et al. [34] extends this work with **DGD2** by comparing MBTI types and temperament theory skill sets (logistical, strategic, tactical, diplomatic). This results in five potential player traits:

[5]Original work presents three components and ten-traits: achievement (*advancement*, *mechanics*, and *competition*), social (*socializing*, *relationships*, and *teamwork*), and immersion (*discovery*, *role-playing*, *customization*, and *escapism*). This was critiqued and expanded on with the Online Gaming Motivations Scale.



PLAYER SEGMENTS SUMMARY

QUANTIC FOUNDRY

	Acrobat	Gardener	Slayer	Skirmisher	Gladiator
Motto	*"Flexing My Reflexes."*	*"Quiet, Relaxing Task Completion."*	*"Cinematic Mayhem With a Purpose."*	*"Jumping Into The Fray of Battle."*	*"Dedicated, hardcore gaming."*
Top Mot.	Challenge + Discovery	Completion	Fantasy + Story + Destruction	Destruction + Competition	Challenge + Completion + Comm,
Pop Games	Spelunky, Celeste, Super Metroid, Tetris	Candy Crush, Solitaire, Animal Crossing	Firewatch, Uncharted, Tomb Raider	Rust, Call of Duty, Battlefield	Mobile Legends, Destiny, Gears of War

	Ninja	Bounty Hunter	Architect	Bard	
Motto	*"A Duel of Speed and Skill."*	*"High-Octane Solo World Exploration."*	*"My Empire Begins With This Village."*	*"Playing a Part in a Grand Story."*	
Top Mot.	Competition + Challenge	Destruction + Fantasy	Strategy + Completion	Design + Community + Fantasy	
Pop Games	Street Fighter, StarCraft, LoL	Mass Effect, Far Cry, Saints Row	Europa Universalis, Civ VI, Banished	The Secret World, FFXIV, LoTRO	

Figure 6.1: Player type summary from Quantic Foundry Gamer Motivation Profile [535]. Figure reproduced from Quantic Foundry Gamer Types reference.

openness to imagination, preference for anger vs. avoidance of frustration, degree of tolerance for real-time play, preference for group vs. solo play, and degree or persistence or obsessive play. Primary work on DGD1 and DGD2 ends here, as Bateman, Lowenhaupt, Nacke, et al. suggests further work on BrainHex.

BrainHex [318] presents seven archetypes representing various configurations of neurobiologically informed traits. The seven types (and some of their MBTIs) are: the curiosity-focused *Seeker* (FP), the fear-oriented *Survivor* (FP), the thrill-seeking *Daredevil* (TP), the strategic *Mastermind* (T), the challenge-seeking *Conqueror* (T), the people-oriented *Socialiser* (F), and goal-motivated *Achiever* (FJ). Further investigations with BrainHex data indicate these seven play styles could not be supported [474]. A three motivation model (*action, esthetic, and goal orientation*) was explored; however, it only had a small-to-moderate effect for predicting player game preferences [474]. Tondello et al. [475] develop a new scale for a five-trait BrainHex model (*aesthetic, narrative, goal, social, and challenge orientation*) which they show is significantly different from the Big 5 model.

Unified Model [451] combines Bartle's Types [30] with Kiersey's Four Tempermants (Artisan, Guardian, Rational, and Idealist) and DGD1 [33] (with additional Hardcore and Casual types) to create a spectrum of player types (Fig. 6.2). This model is "validated" through presenting examples of its usefulness, like a list of common genres with example games and the core play style (i.e. type) they support. Stewart contends that the agreement between these three base models (Bartle, Kiersey Tempermants, DGD1) is "remarkable". However, he glosses over the fact that Kiersey is based on MBTI, and DGD1 explicitly

combines Bartle Types and MBTI - hence clear synergy due to their underlying theory. As well, Stewart is "surprised to see how many of these other models proposed three or four categories...[which] sounded very much like the descriptions of the core play styles in the Keirsey/Bartle model". However, he overlooks validated models with more than four types (e.g. Gamer Motivation Profile [535]) and does not fully address all types in the models he does bring up (e.g. Laws [256]). He implies this supposed pattern reveals there are four fundamental personality types. More likely this is a case of modeling limitations due to data collection (e.g. heterogeneous samples, limited sample sizes, imperfect measurement tools, biased self-reporting), overlap in underlying constructs, or a design choice made to make the model easy to communicate.

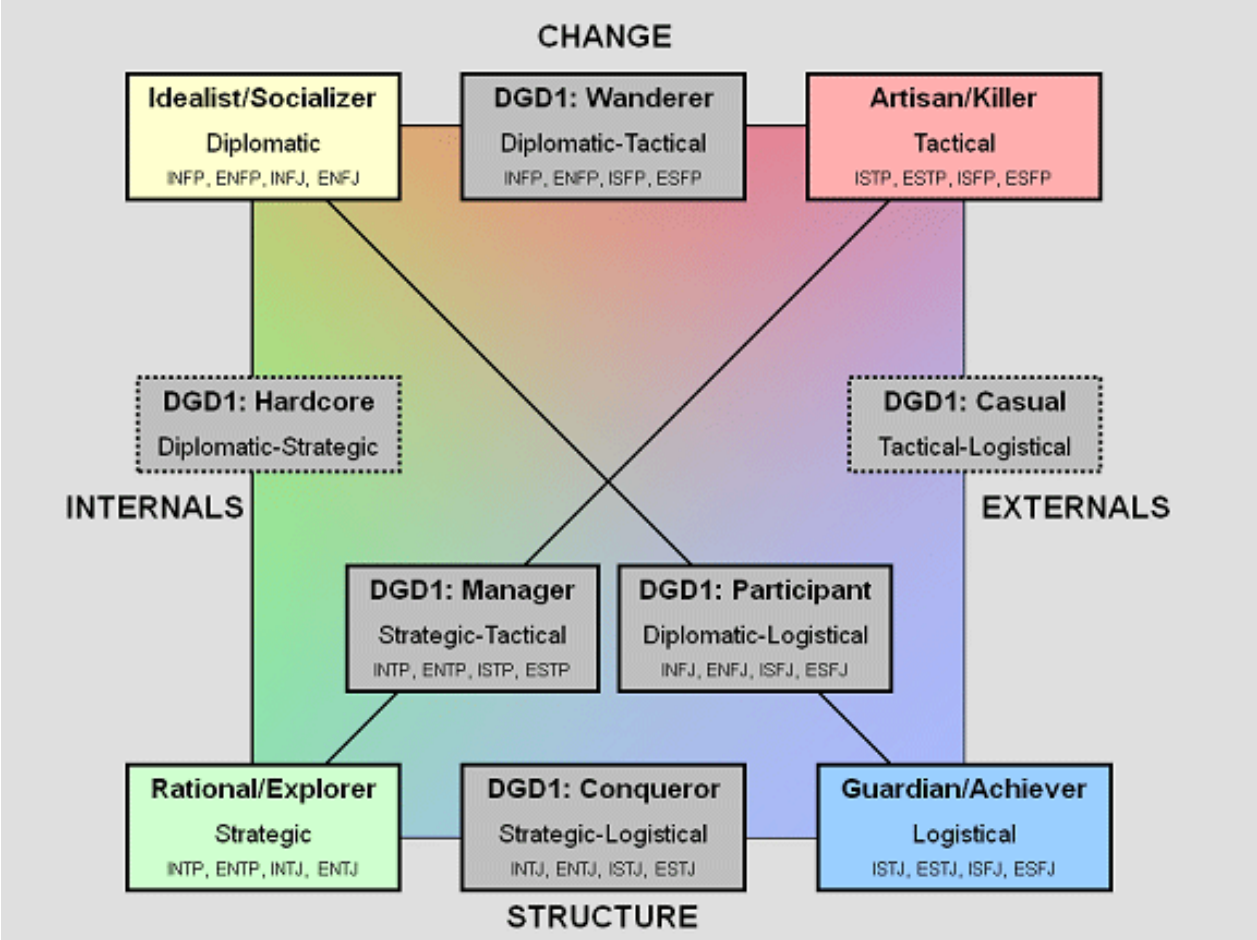


Figure 6.2: Unified Model with DGD1, reproduced from [451].

Tseng [480] proposes two motivations: exploration and aggression. These motivations support three player types: *aggressive gamers* (high exploration and aggression needs), *social gamers* (high exploration, low aggression needs), and *inactive gamers* (low exploration, moderate aggression needs).

Hexad [277] uses self-determination theory (SDT) to match player motivations to potential gamification elements. Hexad presents six player types: Philanthropists (*purpose*), Socialisers (*relatedness*), Free Spirits (*autonomy*), Achievers (*competence*), Players (*extrinsic rewards*), and Disruptors (*change*). The original Hexad scale was found to be unreliable; a second scale was developed and validate [473]. Tondello et al. [473] also validate the Hexad types' relationships to game elements, and compare Hexad types against Big 5 traits.

6.4 Conclusion

Having not found a suitable ready-to-use ability-based model, we are clear in our choice to make our own with the clear goal of evaluating player interactions with challenges through a standard set of cognitive and motor requirements.

Take home points

From this chapter we learned the following meta-lessons:

- Our model should fit into the ability-based design paradigm.
- We like the HCI models' information processing approach and can see it adapting well to our work.
- Existing player models are typically too abstract for our purposes.

Chapter 7

Motor Abilities

We present a succinct review of the motor abilities of our player model. This work originates from Sasha's Masters thesis [442], and is published in Soraine and Carette [441]; we cover it here for completeness. 7.1 explains our scoping decision, construction process and provides a complete motor ability list. The rest of the chapter summarizes the results of our fine and gross motor analysis from Soraine and Carette [441]. For each ability we explain what it is, which actions it combines, and examples of it in commercial games. We end by summarizing key points about the motor model.

7.1 Developing Motor Model

Ch. 4.4's requirements and design decisions scope our motor model to abilities in a current game playing context. We interpret this to mean our model should cover mutually exclusive abilities based on actions possible with commercial game controllers. We consider two abilities to be mutually exclusive when they are different interactions and produce a different motor response. For example, on a Switch Joy-Con the player can both press buttons or shake the controller — different interactions (buttons vs accelerometers) and motor responses (pressing with fingers vs arm-based movements). Our understanding of mutually exclusive abilities captures how the type of motor response affects task performance [6, 27, 370], and makes it easy to identify abilities by the controller inputs. However, we also want our mutually exclusive abilities to "feel" different. Consider two cases on interacting with a standard controller:

Case 1: Face Buttons and Thumbsticks. The interaction methods are technologically distinct (different degrees of freedom, discrete vs analog signal) but the motor responses happen on a similar scale (thumb movements). However, these actions feel different. Pressing face buttons requires thumbs to flex and extend in an up/down motion. The movement is discrete (pressed/not pressed) and does not require precision. Thumbsticks require thumbs to abduct and adduct in a left/right motion. The movement is continuous, and requires precision as small changes in the thumb's motion are captured in the input and reflected in the game performance.

Case 2: Face Buttons and Trigger Buttons. The interaction methods are technologically distinct in quality of information and ergonomics. The motor responses look distinct (pressing face buttons with your thumb vs pulling the trigger with your index finger). However, at the level of action both are pressed in discrete ways and require the same flex/extend motions of the finger to be used. The difference in design here hides that it is the same motor response with the same qualities being used.

As we develop our motor model, we can use qualities like speed, complexity, precision, repetition, and direction to decide whether controller interactions create mutually exclusive abilities. We summarize this in Design Decision (DD) 6.

Design Decision 6. The motor model abilities will capture actions possible on commercial game controllers and be mutually exclusive based on the qualities of the motor response for the interactions.

7.1.1 Scoping the Motor Domains

We initially explore the Keystroke Level Modeling (KLM) of Super Mario Bros 3 [215] since it relates gameplay to controller-based actions. They create two GOMS models (a functional level model, and KLM) of the game from analysing expert behaviour in the first 27 seconds of gameplay (World 1-1), and examining the game mechanics and instruction booklet. While the observational process and mapping of actions to gameplay align with our goals, we recall that GOMS models are inappropriate for our work due the limitations (see Ch. 6.1) and want to focus instead on specific abilities.

We explore kinesiology (Kin) models via undergraduate level textbooks [179, 496]. Actions are modeled by muscle group movements (flexion, extension, abduction, adduction, supination, pronation), such that identical actions use the same group of muscles. This lens is too low-level for our gaming context, as it would consider pressing a button with a curved finger (movement at metacarpophelangeal, proximal interphalangeal, and distal interphalangeal joint) as different than pressing with a flat finger (movement just at metacarpophelangeal joint).

We turn to developmental psychology (DevPsych) textbooks [149, 366, 429] which explores human acquisition and development of motor and cognitive skills. DevPsych models motor abilities through fine and gross skills which describe scale of movement and potential precision. There is no singular set of atomic skills, rather observable tasks are given as skill proxies that can be assessed through motor development scales. This fine-gross skill paradigm seems reasonable for our motor model, and its task dependent presentation means we can define a standard set for a gaming context based on surveying controllers. The DevPsych Fine-Gross paradigm also works with the GOMS style approach we like since both come from a place of task analysis via ability descriptions.

7.1.2 Motor Model Method

We explain our three step model development process below, and summarize the major results. Details can be found in Soraine [442].

1. Controller Survey. We focus on standard controllers (e.g., Xbox One controller, Playstation 4 controller), handheld motion controllers (e.g., Wii Remote, Playstation Move), full body motion controllers (e.g. Kinect), smartphones/tablets, handheld consoles (e.g. Nintendo 3DS, Playstation Vita, Nintendo Switch), keyboards, mice, and specialty controllers like fight sticks (arcade style controllers made for fighting games), mat controllers (e.g. DDR dance pad, Wii Balance Board), and "simulation" contollers (thematic controllers like Donkey Konga Drums). For each controller type we list its interactive hardware (e.g. face buttons, triggers, thumbsticks, gyroscopes, etc) and a natural language description of the interaction (e.g. pressing face button, pulling trigger button). We group similarly described interactions to minimize redundancy; for example, pressing a face button and pressing a D-Pad would come together as examples of a more abstract "pressing" action.

2. Mutual Exclusivity. We label actions as either Fine or Gross, and organize them based on the underlying motor system. For example, standard controllers, handheld consoles, and mat controllers all afford "pressing" buttons. However, the standard controller and handheld console buttons are Fine motor skills using Hands/Fingers, while the mat controller is a Gross motor skill from stepping on it with your Legs/Feet. We summarize the abstract actions, their hardware contexts, and general category (Fine or Gross motor) in Table 7.1.

Actions	**Hardware Context**
Pressing	SC, HMC, HC, Key, FS, MC
Bumping	SC, HC
Pulling	SC, HMC
Moving	SC, HMC, FBM, HC, FS, M
Swiping	SC, Phone
Pinch-to-zoom	SC, Phone
Swinging	HMC
Pointing	HMC
Shaking	HMC, Phone, HC
Drawing	HMC, Phone, HC
Thrusting	HMC
Tilting	HMC, Phone, HC
Flicking	HMC, Phone
Positioning	FBM, Mat
Tapping	Phone, HC
Speaking	Phone, HC
Making facial expressions	HC
Clicking	M
Scrolling	M

Legend: Standard Controllers (SC), Handheld Motion Controllers (HMC), Full Body Motion Controllers (FBM), Smartphones/ Tablets (Phone), Handheld Consoles (HC), Keyboards (Key), Mouse (M), Fight Sticks (FS), Mat Controllers (Mat)

Table 7.1: Resulting actions from controller survey, reproduced from [442]. Blue rows are Fine motor, yellow rows are Gross motor, green rows are abilities in both.

3. "Feel" Different. We analyse the qualities of each action via examples, focusing on speed, complexity, precision, repetition or direction of motion between abilities. We combine actions that are not sufficiently different, and rename the group based on the most easy to visualize action.

Using this process we develop a model of 13 Fine and 8 Gross motor abilities based on controller interactions (Table 7.2). Our model is biased towards Fine motor abilities because controllers generally favour hand-based interactions (see Table 7.1). Our process produces a working list that can be expanded when new controllers are analysed. We believe this list provides reasonable ability coverage based on the variety of controllers examined; however, it is not comprehensive.

Category	Motor System	Subsystem	Ability
Fine	Hands	Fingers	Pressing
			Swiping
			Pinching
			Shaking
		Wrist	Flicking
			Pointing
			Swinging
			Drawing
			Tilting
	Head	Neck	Moving
		Face	Speaking
			Making facial expressions
	Feet	Ankle and Foot	Pressing
Gross	Arms		Pushing
			Swinging
			Drawing
			Rotating
			Positioning
	Legs		Moving
			Positioning
	Torso		Positioning

Table 7.2: Final list of motor abilities for the player model. Adapted from [442].

7.2 Fine Motor Abilities

We present the results of our fine motor analysis from Soraine and Carette [441]. We organize the abilities by their motor sub-system from Table 7.2. We find more Fine Motor abilities because controllers tend to prioritise hand-based actions, leading to manual dexterity being important in game performance.

7.2.1 Fingers

Pressing: We combine *pressing*, *clicking*, *tapping*, *pulling*, and *bumping*. Each action describes moving or bending a finger at a knuckle for the purpose of interacting with a control. Their main difference is their associated control element: face button for pressing, mouse button for clicking, touchscreen for tapping, trigger button for pulling, shoulder button for bumping. As well, the hardware context has pulling and bumping oriented differently because of the interaction position on the controller. These differences lead to minor experiential differences in feedback, but overall the "feels" seem the same. As pressing buttons are ubiquitous in games we do not give a specific example.

Swiping: We combine *swiping*, *flicking*, and *scrolling*[1]. Each action describes fluid, potentially repetitive finger movements; however, they differ in time. Conceptually flicking is a rapid movement, where swiping and scrolling can be fast or slow. At the moment we are coarsely defining time differences in action as changing their "feel". All three actions take place in the same time scale (see Newell time band [325]), meaning we consider their differences negligible.

Pinching: Coordinated two-finger movement to create a pincer-grip/pinching motion on a touch-sensitive surface (e.g. smartphones/tablets, handheld consoles). Pinching represents a single-task coordinated action (STCA), which has been shown to increase cognitive load for older adults (e.g. 159, 266, 273, 360, 425). Therefore since it is measurably different from other actions, we keep it separate.

7.2.2 Wrist

Pointing: Continuous precise lateral (Fig. 7.1a, like waving as a greeting) and vertical (Fig. 7.1b, like fanning oneself) wrist movement. Common in pointing tasks, like Super Mario Galaxy [339].

(a) Lateral Wrist Flexion/Extension. (b) Vertical Wrist Deviation.

Figure 7.1: Wrist pointing movements with a Handheld Motion Controller (Wii Remote).

Flicking: Fast, imprecise, short, singular lateral wrist movement. Flicking is a supporting motion in many challenges like serving the ball in table tennis for Wii Sports Resort [336].

[1]Scrolling only refers to scroll-wheel interactions, as touchscreen "scrolling" is a general case of Swiping.

Tilting: Continuous coordinated wrist and forearm movements/rotations with contextual differences based on hold. For one-handed holds, tilting laterally involves the player twisting their wrist and forearm as if turning a doorknob (Fig. 7.2a). Tilting vertically in this context is the same movement as vertical pointing movements (radial and ulnar deviation). For two-handed holds, tilting laterally requires the player's forearms guide the movement while the wrist keep the controller stable (Fig. 7.2b). Tilting vertically uses the same movements as vertical pointing (radial and ulnar deviation). Examples include: steering the flying beetle in Legend of Zelda: Skyward Sword [332] (one-hand vertical and lateral), balancing on skateboard in Mario and Sonic at the Tokyo 2020 Olympic Games [419] (one-hand lateral), and the Myahm Agana Shrine in Legend of Zelda: Breath of the Wild [344, 345] (two-hand vertical and lateral).

(a) One-Handed: Wrist Supination and Pronation.

(b) Two-Handed: Forearm Flexion and Extension.

Figure 7.2: Lateral Wrist Tilting. One-Handed with a Handheld Motion Controller (Wii Remote); Two-Handed with a Handheld Console (Wii U).

Drawing: Controlled continuous wrist/forearm movements with contextual differences based on the "canvas" (i.e. control space). A small "canvas" (e.g. smartphone touchscreen) employs wrist movements; a large "canvas" (e.g. the air) enables forearm movements. Examples include the Celestial Brush mechanic in Okami! (particularly on the Wii) [90], and Kirby: Canvas Curse [175].

Swinging: Repeated (back and forth) large lateral wrist movements. Flicking and swinging are distinct because of their different speeds (flicking must be fast) and number of movements (swinging needs a back and forth). Examples include: using the fishing rod and net in Animal Crossing: City Folk [335], cracking an egg in Cooking Mama: Cook Off [96], and sword actions in The Legend of Zelda: Skyward Sword [332].

Shaking: Fast, imprecise, short, repetitive wrist/forearm movements contextualized by the controller hold. One-handed holds mimic a drumstick tapping on a drum, or as a jerking forearm movement similar to the motion of shaking a cocktail shaker (Fig. 7.3). Two-handed holds use the same forearm movements as two-handed tilting. Examples include: ground pound using Wii Remote in Donkey Kong Country Returns [390] and Tropical Freeze [391], asteroids in SpaceTeam [436], wheelies in Mario Kart 8 [328], and the homing hat throw in Super Mario Odyssey [343].

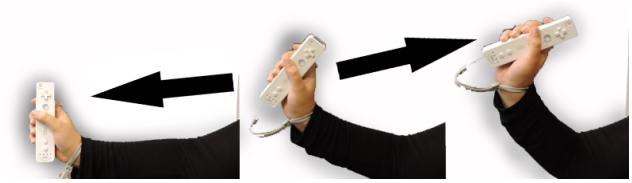


Figure 7.3: Forearm Shaking (forearm rotation) with a Handheld Motion Controller (Wii Remote).

7.2.3 Head: Neck and Face

Neck Moving: We combine head *tilting*, *nodding*, and *shaking* as neck movements. These actions are becoming more important for social AR games, like Facebook's Asteroids Attack [131] and VR games, which use head movements for camera controls.

Facial Expressions: Controlled coordinated movement of facial features. This can be used for social gameplay, like Pokémon Amie in Pokémon X and Y [150, 151], or as a core mechanics like in Snapchat's Snappables [440].

Face Speaking: Making controlled noise directed at the control's microphone. This is not natural language processing, as it only concerns amount of noise and intensity of noise. Examples include Puzzle 138 in Professor Layton and the Diabolical Box [263], which requires players to blow into their microphone simulating a gust of wind, and Chicken Scream [373] on smartphones, which allows the user to control how the chicken avatar moves by making sounds.

7.2.4 Ankle and Foot

Pressing: Coordinated movement of the ankle and foot to depress a button. While there may be other ankle/foot movements, current foot-focused controllers (i.e. mat controllers) are limited in their use. Examples include Dance Dance Revolution [237], Shaun White Skateboarding [487], and Mario and Sonic at the Winter Olympic Games for the Wii Balance Board [422].

7.3 Gross Motor Abilities

We present the results of our gross motor analysis from Soraine and Carette [441], organized by their motor systems. These abilities are significantly broader than their Fine motor counterparts due to the controllers surveyed. This list was made circa 2018, meaning more recent motion controls are not captured.

7.3.1 Arms

While arms have multiple parts (upper arm, forearm, elbows and shoulders) we treat them like one unit in our model. This is an oversimplification based on the controllers surveyed

and the time this list was compiled.

Pushing: We combine *moving*, *pressing*, *thrusting*, and *pulling*. These describe large coordinated shoulder and elbow movements (flexion and extension) to interact with a physical element (e.g. pushing a large button, pushing a mouse on a table) or a simulated element (e.g. pushing a button in VR, punching in Wii Sports Boxing [333]). These movements can be fast and imprecise or slow and controlled, as our currently studied gaming context does not see these motor responses as "feeling" different[2].

Rotating: Controlled medial and lateral shoulder rotations used for moving controls laterally (e.g. moving mouse left and right) and simulating big motions like Golfing in Switch Sports [341].

Swinging: We combine *swinging*, *flicking*, and *shaking*. These are fast repetitive context-dependent shoulder and elbow movements. Vertical swinging (like chopping wood) uses a combination of shoulder and elbow flexion/extension, while horizontal swinging uses medial and lateral shoulder rotations. Swinging can be either precise (e.g. sword moves in Legend of Zelda: Skyward Sword [332], Chambara in Switch Sports [341]), or imprecise (e.g. Swimming and Discus Throw in Mario and Sonic at the Tokyo 2020 Olympics [419]).

Drawing: Precise coordinated arm movements that can be fast or slow. Examples include tracing lines, shapes, and stars in the Trauma Center series for Wii (Second Opinion [18], New Blood [17], and Trauma Team [19]) — see Fig. 7.4.

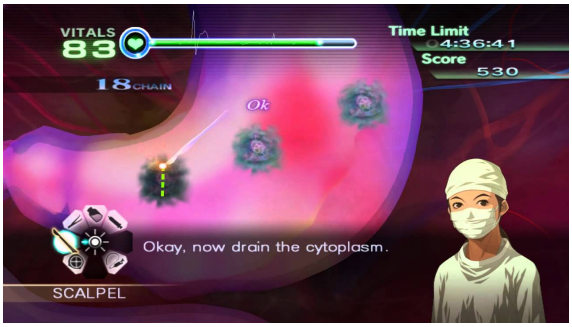


Figure 7.4: Video clip of Trauma Center New Blood gameplay from Youtuber roguehwz. Video shows player in operation having to trace lines to perform actions. Click image to play.

Positioning: We combine *positioning* and *pointing* to describe manipulating our arms into a specific position. This is common in games like Just Dance![488], where players mimicking the movements of on screen dancers.

[2]This will be different given a modern review of VR gaming

7.3.2 Legs

Similar to arms, we oversimplify legs by treating them like a single unit due to the limited leg-based controls at the time of our survey.

Positioning: Manipulating legs into particular positions. Examples include Wii Fit's yoga games[337], and Kinect Adventures! 20, 000 leaks mini-game [161] (see Fig 7.5).

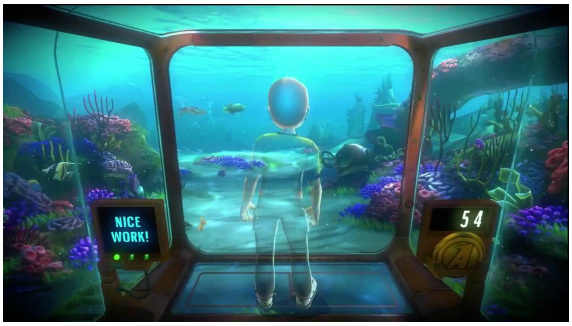


Figure 7.5: Video clip of 20, 000 Leaks! gameplay from Youtuber ChuckieJ Gaming. Video shows gameplay of player's avatar (controlled via Kinect) positioning their body to block leaks. Click image to play.

Moving: We combine all potential movements into one *moving* category. We present examples of specific types of movement in Table 7.3. At the time of creating this list, leg-based interactions were limited by the commercial technology. Full body motion controls like Kinect required very specific conditions to effectively track the player. Mat controllers reduced the leg movements to a button press making the leg generally unobserved. As well, gameplay was limited to exer-games, rhythm games, and hardware showcasing games. Since this time, more leg-based actions have been made possible through exergames like Ring Fit Adventure [340] and its handheld motion controller peripherals which allow for tracking squatting, walking, thigh squeezing and more.

Action	Control	Example
Kicking	FBM	Soccer (Kinect Sports [387])
Jumping	FBM	River Rush (Kinect Adventures [161])
	Mat	Jumps[3] (Dance Dance Revolution [237])
Stepping	FBM	River Rush (Kinect Adventures [161])
	Mat	Single step (Dance Dance Revolution [237])
Jogging	FBM	Sprint (Track and Field) in Kinect Sports [387]

Table 7.3: Leg actions with in game examples

[3]Simultaneous two button press

7.3.3 Torso

Positioning: Bending and twisting the torso to support other movements and positioning. These types of movements are usually found in dance games, like Dance Central [181], and simulation games like Yoga in Wii Fit Plus [338].

7.4 Wrapping up the Motor Model

While our proposed motor model is somewhat oversimplified and limited based on when it was made, we do not think this will be a problem. As a first foray into this domain, it is sufficient to focus on gameplay in fairly standard and simple contexts (e.g. standard controllers, keyboards). So our Fine Motor Abilities are likely sufficient for our thesis goals. Future work should focus on incorporating new controllers, especially as it relates to more clarity in gross motor abilities.

Take home points

From this chapter we learned the following meta-lessons:

- Our motor model is built from a survey of controllers and actions used in games.
- We attempt to make the motor abilities mutually exclusive based on the qualities of the motor interactions.
- Our final motor model has 21 abilities (Tbl. 7.2) — broken into 13 Fine and 8 Gross motor abilities.

Chapter 8

Cognitive Abilities

This chapter outlines the cognitive component of our player model. We present our theoretical framework, scope, and construction method in 8.1, along with a complete cognitive ability list. The rest of the chapter explains the abilities, organized by cognitive system, and how they manifest in existing games.

8.1 Developing Cognitive Model

We use the information-processing model (InfoProc, Fig. 8.1) to understand cognition. We focus on *Perceptual Encoding* to *Central Processing* stages[1]. Cognition involves three independent subsystems (perception, attention, memory) which interpret and process information [518, p. 146]. We need to find the specific cognitive processes inside these subsystems.

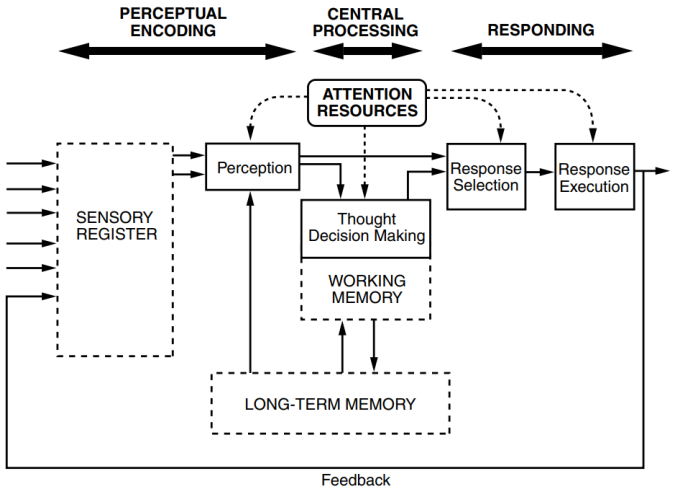


Figure 8.1: Generic Information Processing Model reproduced from Wickens et al. [518, p. 147]

[1] *Responding* is covered by our Motor Model (Ch. 7).

Scoping: We look to cognitive psychology (CogPsych), Human-Computer Interaction (HCI), and developmental psychology (DevPsych) to build our broad understanding of the processes (i.e. abilities) of each subsystems. We include DevPsych because its focus on normative development and ability acquisition helps us understand the expected abilities that a neuro-typical individual will have at different points in human development.

Method: We survey *Cognitive Psychology, 6th edition* [129], *Introduction to Human Factors Engineering, 2nd edition* [518], *Psychology: Themes and Variations, 6th edition* [515], and *Cognitive Development: Infancy through Adolescence* [149] to learn about the subsystems. We supplement this base with citation searching and searching for specific abilities via GoogleScholar. We scope the resulting abilities to a gaming context[2] based on the game's communication modalities: audio, video, and haptic. This mainly affects the perception processes in our model. To show how each ability fits into gaming, we provide examples of them in gameplay. This results in 25 abilities summarized in Table 8.1.

System	Subsystem	Ability	Textbook
Perception	Vision	Facial Recognition	129, 515
		Object Recognition	129, 515, 518
		Token change detection	129, 515
		Type change detection	129, 515
		Heading	129, 515
		Steering	129, 515
		Time to Contact	129, 515
	Audial		129, 515
	Tactile		129, 515
Attention	Selective Attention	Auditory	129, 515
		Visual	129, 515
	Divided Attention		129, 515
Memory	Sensory Stores	Echoic	129, 515
		Iconic	129, 515
	Long Term Memory	Episodic Memory	129, 515, 518
		Semantic Memory	129, 515, 518
		Perceptual Representation	129, 515, 518
		Procedural Representation	129, 515, 518
	Working Memory	Phonological Loop	129, 515, 518
		Visuo-spatial Sketchpad	129, 515, 518
		Episodic Buffer	129, 515
		Inhibition	129, 515
		Task Shifting	129, 515
		Updating	129, 515
		Multi-tasking	129, 515

Table 8.1: List of cognitive abilities found through literature review.

[2]Per Ch. 4.4 requirements.

A Quick Aside...

To help show the abilities in gameplay, framed images in this chapter are clickable links to YouTube videos. Captions reference the particular game and the YouTube creator.

8.2 Perception

This system *interprets* sensory information, hence it works closely with memory. Perception abilities are separated by information modality (audial, visual, tactile). We focus on understanding *visual perception* since it is the main communication mode in games. Visual perception has two purposes: recognition (vision for perception) and action planning (vision for action) [162, 173, 209, 210, 300–303, 352, 372].

8.2.1 Vision for Perception

Underlying processes are object recognition, facial recognition, token change detection (replacing an object with one of the same type), and type change detection (replacing an object with one of a different type) [43, 120, 129, 133–135, 153, 189, 191, 278, 279, 290, 291, 295, 306, 400, 408, 465, 478, 538].

Figure 8.2: Senseless Census (Super Mario Party) by arronmunroe.

Object recognition is a combination of recognition by feature composition with semantic memory. It appears in visual search style gameplay like the Abandoned House level in Donut County [38]; and mini-games like Absent Minded (Super Mario Party [324]) during levels where you have to identify pixelated characters. It usually occurs alongside choice reaction tasks (e.g. Looking for Love, and Sorting Fun from Super Mario Party), and it can often be a source of difficulty in games which ask you to tell the difference between similar looking objects (e.g. Senseless Census - Fig. 8.2).

Facial recognition covers recognizing faces, facial expressions and their associated emotion. L.A. Noire [466] interrogations rely on this as players deduce whether an NPC is lying from their facial expressions (as noted in the official strategy guides [48]). Facial recognition is also used in Making Faces (Super Mario Party [324]), where the player has to place parts of a face on a blank canvas to accurately recreate either Mario or Bowser (Fig. 8.3).

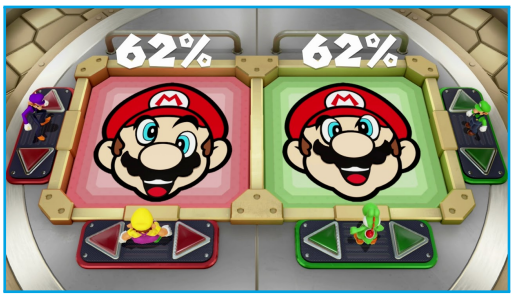


Figure 8.3: Making Faces (Super Mario Party) by arronmunroe.

Change detection (token and type) [105, 188, 190] combines object recognition with attention.

Figure 8.4: TCD used to recognize bomb-able wall spots.

Token change detection (TCD) is when an object is replaced by another object of the same type. It occurs in exploration challenges, like recognizing a bomb-able wall in Legend of Zelda: A Link to the Past [330] (Fig. 8.4), or as puzzles/logic challenges like find paperwork discrepancies in Papers Please [1]. **Type change detection (TyCD)** is when an object is replaced by another object of a different type, like swapping out a Pikachu for a Charmander in a Pokémon battle. It is used in spot-the-difference challenges (e.g. Spot the Addtion in Fable Anniversary [267]).

8.2.2 Vision for Action

Underlying processes are heading (direction orienting), steering (direction movement), and time to contact [39, 64, 129, 138, 174, 245, 258, 397, 410, 411, 470, 521, 522, 537, 541].

Heading and Steering respectively cover orienting and moving characters/avatars through the game space. These are the basic abilities used in exploration challenges, like Slender: The Eight Pages [363], or the dream segments of Try to Fall Asleep [10]. Games with diegetic maps, like Sea of Thieves [388] (Fig. 8.5) and Firewatch [75], rely more heavily on these skills as players have no meta-game support. Racing game series like Mario Kart and Forza also employ these abilities using the car in place of the player avatar.

Figure 8.5: Heading/Steering in Sea of Thieves [388] using diegetic map to navigate the island.

Figure 8.6: Night Light Fright (Mario Party Superstars) by Paranoia's Dungeon.

Time to contact allows us to estimate the distance between ourselves and moving objects. In real life we use it to help catch thrown objects, or avoid incoming objects. We use it similarly in games like Night Light Fright (Mario Party Superstars [323], Fig. 8.6) where we try to stop an approaching Chain Chomp when it is as close to us as possible, but not touching us. Time to contact also factors into reaction time and rhythmic challenges, like All-Star Swings (Super Mario Party [324]) where the player tracks an incoming baseball to be hit at the right moment.

8.3 Attention

Under a processing model, attention is a set of resources used to engage focus on stimuli [52, 129, 165, 184, 221, 351, 384]. We subscribe to the *multi-store resource pool model* which outlines different resource pools for each sense [165, 221, 351]. In our gaming context we care about auditory and visual resources. These resources are handled by two processes: selective and divided attention [65, 88, 129, 254, 305, 384, 476].

Selective attention allows us to target our focus on a specific stimulus by consuming attentional resources. We can focus on multiple stimuli if they pull from different resource pools, as we have different limitations depending on stimulus modality [80, 211]. Selective Attention is important across gameplay. It features in combat mechanics as players must focus on their opponent's animations and actions to respond appropriately. It plays a role in pattern and rhythm challenges, like Snore War (Pokemon Stadium [334]), where players must watch a giant swinging pendulum and hit the A button exactly when the pendulum reaches its equilibrium position to hypnotize their opponents (Fig. 8.7). It is key to reaction time challenges where it works alongside time to contact to focus on the object in motion. An example is Rock Harden (Pokemon Stadium [334]) where players must hit the A button at the right moment to avoid taking damage from a falling rock.

Figure 8.7: Snore War (Pokemon Stadium) by Hawlo.

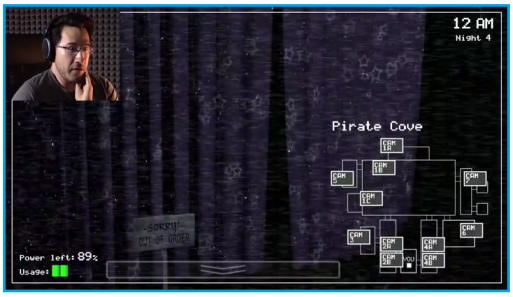


Figure 8.8: Night 4 of Five Night's at Freddy's by Markiplier.

Divided attention allows us to spread concentration between multiple stimuli of the same type for the purpose of completing a singular task. It is frequently a passive challenge in larger games like League of Legends [395] and Resident Evil 7 [77], where players are watching for enemies around the environment. It is also used alongside multi-tasking (a Working Memory ability, see 8.4.2) when players must monitor multiple resources while performing other actions. A simple example is Five Nights At Freddy's [418] (Fig. 8.8), where players must survive for a set amount of time while being accosted by four animatronics. The player must spread their attention across in-game monitors, and doorways to track the animatronic's locations all while keeping an eye on the limited supply of power they have which is consumed by checking the monitors and doors.

8.4 Memory

Memory handles information encoding and retrieval. We understand it as a multi-store model which separates memory into a sensory store, working memory (WM), and a long-term memory (LTM) [24, 25].

8.4.1 Sensory Stores

Sensory information is temporarily held by *echoic* and *iconic* sensory stores [129, 251, 443, 477]. These stores are constantly engaged, taking in sensory stimuli and feeding it to the perception system. The information only lasts from milliseconds to seconds.

8.4.2 Working Memory

WM holds information for short term processing and thinking. We use Baddeley's model [24], which highlights four subsystems: the phonological loop, visuo-spatial sketchpad, central executive, and episodic buffer[21, 24–26, 98, 99, 129, 253, 308, 312, 482].

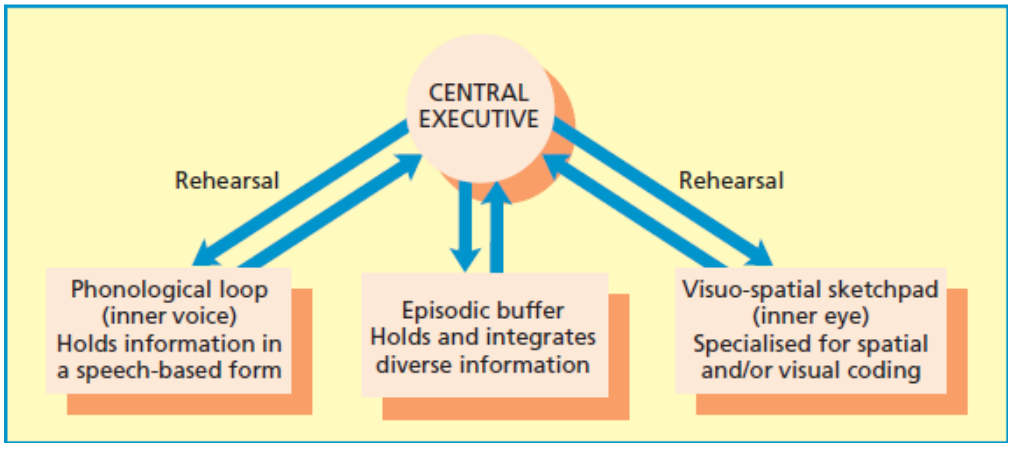


Figure 8.9: Baddeley's Model of Working Memory [24], reproduced from *Cognitive Psychology, 6th edition*[129, p. 212]

The phonological loop handles the processing and creation of sounds. It has two parts: the phonological store (remembers words we hear) and articulatory process (rehearses and repeats words) [22, 129, 186, 269, 272, 479, 490]. These parts are essential in communicating with other players. It is the main challenge in social games like Spaceteam! [436] (Fig. 8.10) where players have to communicate non-sensical instructions to keep their space ship going.

Figure 8.10: Spaceteam by TeamHypercube.

Figure 8.11: Memory Match (Mario Party) by NintendoMovies.

The Visuo-spatial sketchpad remembers and processes object and spatial information. It is composed of the visual cache (remembers object shape and colours), and the inner scribe (remembers object movement and spatial layout, and transfers information to the central executive) [129, 234, 404, 438]. The visuo-spatial sketchpad is generally needed to interact with game spaces, and is the primary ability for matching games, like Memory Match (Mario Party [193], Fig. 8.11).

The episodic buffer integrates information from the phonological loop and visuo-spatial sketchpad and transfers it to long term memory [23, 42, 129, 163, 401]. It is useful for developing game knowledge over time.

The central executive has four major processes: inhibition, task shifting, updating the working memory state, and multi-tasking [21, 92, 129, 144, 271, 304, 308, 454].

Figure 8.12: Fruit Ninja gameplay by Neogaming.

Inhibition stops us from responding to certain stimuli. It is a central element in many games; one example is Fruit Ninja [177] where the player must avoid cutting the bombs (Fig. 8.12).

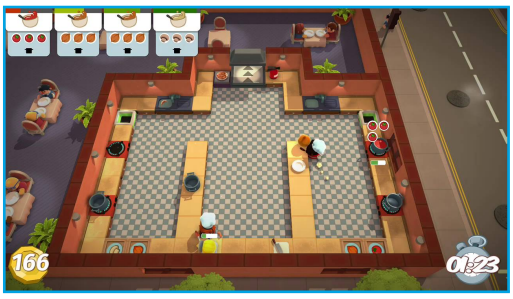


Figure 8.13: Overcooked (Co-op) by Colin Kelly.

Task shifting refocuses attention between different tasks. It is used frequently as players move between different activities during their gameplay sessions. For example Overcooked! [155] requires players to switch between preparing, cooking, running and cleaning up after food. While they switch between these tasks they need to watch the progress of the other tasks so nothing burns or piles up (Fig. 8.13). In this way, task shifting commonly occurs alongside divided attention and multi-tasking to create difficulty.

Updating adjusts our working mental models based on new stimuli and information. It is used alongside other abilities in games like Suit Yourselves [324], where the player is shown the location of cards on a board and then the board is rotated (changing the card position) before asking the player to find a specific suit of card.

Figure 8.14: Suit Yourselves (Super Mario Party) by NintendoMovies.

Multi-tasking is the process of simultaneously doing tasks that have different goals. It is used across all games; we have already seen examples of this with our examples of Five Night's at Freddy's (Fig. 8.8) and Overcooked (Fig. 8.13).

8.4.3 Long Term Memory

LTM stores two types of learned information: declarative and non-declarative[129, 416, 417, 482]. LTM is built from our experience playing games, and is the foundation of our game literacy. We use it whenever we interact with a game (as we remember the controls), and it enables us to engage with more complex processes like problem-solving. While these higher order processes are out of scope of our thesis, we do reference them to contextualize our LTM game examples.

Declarative memory has two subtypes: episodic (personal events and specific episodes), and semantic (general knowledge about the world, concepts, language, etc.)[129, 166, 224, 307, 364, 382, 444, 481, 483, 497, 498, 507, 516]. **Semantic memory** encodes and retrieves learned information. For playing video games it is used to learn information about mechanics, controls, etc. **Episodic memory** is an ability that encodes and retrieves memories of events and experiences. In games this helps you remember the plots of the games, and experiences you had during gameplay and between sessions.

Semantic and episodic memory work together to build strategies and problem solve in games. Consider you are playing Final Fantasy X [447] and encounter a Bomb enemy. After five moves, the Bomb explodes, damaging your party but killing itself. The next time you encounter a Bomb you remember this fact (episodic memory) and know that it is coming. So you wait to see if it happens after 5 moves again. It does, so now you know that Bomb enemies explode after 5 moves. This cycle of experience and testing creates information that then can be encoded into semantic memory about the Bomb enemy which persists between games. So when you encounter a Bomb in Final Fantasy XII [446] you remember it explodes after 5 moves, which guides how you approach the fight.

Non-declarative memory has two types: perceptual representation and procedural representations [85, 129, 148, 167, 264, 379, 380, 414, 444, 520]. **Perceptual representation** is the memory side of recognition; it encodes and retrieves information about the different stimuli being perceived. It also allows for faster recognition of stimuli that has been previously encountered and is considered responsible for the priming effect. In games this helps in the recognition of important items. **Procedural representation** encodes and retrieves information about how to perform tasks and actions. In games when you practice a sequence of controls to pull off a combo in a fighting game, you are relying on procedural representation of the move.

8.5 Wrapping Up the Cognitive Model

Our current model focuses on very basic cognitive systems, and very functional processes. While this is sufficient for our scope, it does inherently limit the types of games we can look

at with this model. Future work should explore how to expand into higher-order processes for reasoning and problem-solving. Those would enable us to model more abstract types of gameplay, and they could be useful in modeling how player's interact with multiple challenges at once. It is possible that the existing processes are sufficient, as the central executive has been found to play a large role in task planning, initiation, organization, and monitoring [129, p. 218][3]. However, we would need to explore these abilities in more detail.

Take home points

From this chapter we learned the following meta-lessons:

- We use the information processing model to guide our work, and view cognition as three independent systems: perception, attention, and memory.
- We find 25 cognitive abilities (Tbl. 8.1) relevant to a gaming context.
- The model is currently geared towards basic, functional processes and would need more work to address more involved cognitive tasks.

[3]Notably studied through impairments of executive function.

Chapter 9

Measuring Player Abilities

Recall we aim to model players and challenges in terms of cognitive and motor abilities. By combining the motor (Ch. 7) and cognitive models (Ch. 8) we create our full ability set scoped to a gaming context. We now start working towards measuring these abilities (the actual modeling of our Player Model).

Cognitive psychology (CogPsych) measures abilities through carefully designed tasks. We explore literature on existing tasks for our abilities, paying specific attention to their designs, metrics, and administration. From this search we aim to compile a list of tasks that we could administer to players to construct their player profiles.

We begin by scoping our search to the abilities used in Button Mashing challenges. From there we survey measurement methods from CogPsych and present an overview of tasks for each ability. This list will need to be further refined for actual implementation, but that process is covered in Ch. 10 for readability.

9.1 Scoping Ability Measurement Survey

We have 46 unique abilities (21 motor, and 25 cognitive), which makes it infeasible to explore and implement tests for all of them in this singular thesis. Recall our larger goal is modeling mechanical experiences of gameplay challenges through these abilities. So we choose to scope our player profiling to the subset of abilities used in the challenge we will model. We focus on *button mashing challenges* as the simplest atomic challenge, and therefore base test case for our model[1]. From Ch. 14, we identify the following abilities as part of button mashing, and therefore our current ability search scope:

1. finger pressing,
2. wrist pointing,
3. selective attention,
4. inhibition,
5. object recognition,
6. token change detection,
7. tactile perception, and
8. procedural memory.

[1]Full scoping rationale for button mashing challenges can be found in Ch. 13.3.

We conduct a broad search for standardized measures with the goal of understanding their structure, metrics, and design decisions. We start with *Computerized Assessment in Neuropsychology: A review of tests and test batteries* [222], and *A Compendium of Neuropsychological Tests* [82]. We supplement this base with experimentally-based literature found through searching for the abilities and tests in PsycInfo (via ProQuest) and Google Scholar. We report 24 unique test types (Table 9.1). Not all of these will be used in our model; Ch. 10.2 covers our scoping to a subset.

Ability	**Tests**	**Metric**
Finger pressing	Finger tapping test	No(P)/10s
Wrist pointing	Steadiness Test	No(probe touches side)
Selective attention & Inhibition	Simple GNG	RT_{ms}, No(C), No(I), ErrT
	Parametric GNG	RT_{ms}, No(C), No(I), ErrT
	AX-CPT	RT_{ms}, No(C), No(I), ErrT
	Flanker CPT	RT_{ms}, No(C), No(I), ErrT
	TOVA	No(C), No(I), ErrT
	Stroop Colour-Word Test	RT_{ms}, No(C), No(I), ErrT
Object recognition	Object detection	No(C)
	Object Decision	No(C)
	Object Categorization (broad)	No(C), No(I), Acc
	Object Identification	No(C), RT_{ms}
	Shape Detection	No(C)
	Incomplete Letters	No(C)
	Silhouettes	No(C)
	Progressive Silhouettes	No(C)
	Novel object recognition	No(C), No(I)
Token change detection	Canonical CDT	No(C)
	Partial Report CDT	No(C)
	Multiple Change CDT	No(C)
Tactile perception	Haptic ORT	No(C), No(I), Acc, TT(feel objects)
Procedural memory	PRT	ErrS, Error timestamps
	SRTT	RT_{ms}, No(Err), ErrT
	MTT	TT(Complete test), No(Err), ErrS

Tests: Go/No-go task (GNG), Continuous Performance task (CPT), Test of variables of attention (TOVA), Change detection test (CDT), Object recognition tasks (ORT), Pursuit rotor task (PRT), Serial reaction time (SRTT), Mirror tracing task (MTT)

Metrics: Number of X (No(X)), Presses (P), Correct responses (C), Incorrect responses (I), Accuracy (Acc), Errors (Err)[2], Types of Errors (ErrT), Size of Errors (ErrS) Total time to X (TT(X)), Response time in X (RT_X), Seconds (s), Milliseconds (ms)

Table 9.1: List of tests for measuring the abilities used in button mashing challenges.

[2]Errors are treated separately from incorrect responses because it covers all types of errors.

9.2 Survey of Ability Measurement Methods

We organize the tests by their measured ability. For easy reference, we hyperlink the section titles (i.e. ability names) to their definitions. We note that our motor abilities are not standardized, and so there is a possibility that we cannot find specific, formalised, and validated tests for them. In these cases, we report on tests that seem to approximate the ability. When describing these tests, we refer to the person being assessed as the "testee" for simplicity.

9.2.1 Finger Pressing

Finger Tapping Test (FTT) (e.g. 32) has the testee press a button with their right-hand index finger as quickly as possible for ten seconds. FTT measures motor coordination as the average number of presses across five trials. It takes about one minute to complete per finger being tested. FTT is predominantly analog; testees are told when to start and stop pressing, and they use specialized measurement equipment (a box with a single button). Computerized versions substitute the box with a keyboard or mouse, and visually indicate when to start and end tapping. Computerized versions vary on whether miss-clicks are counted as errors, or whether they invalidate the trial.

9.2.2 Wrist Pointing

There are no tests for this ability. We explore broad motor coordination tests, like the *Trail Making Test*, *Grooved Pegboard Test*, and *Purdue Pegboard Tests*. These focus on studying manual dexterity, but from their descriptions rely on multiple wrist abilities. The closest approximation to wrist pointing, as we understand it, is the *Steadiness Test*.

Steadiness Test (e.g. 192, 435) is similar to the children's game Operation [216]. Testees place a probe into holes of various diameter on a steel sheet. They must hold the probe in the hole without touching the sides for a specified amount of time (90 seconds in Simon [435]; 20 seconds in Hudgens et al. [192]). Touching the sides results in a buzzer noise. The tester records how frequently and for how long the probe touched the side (based on buzzer). The specialized equipment makes this test exclusively analog.

9.2.3 Selective Attention and Inhibition

These abilities always appear together and so are measured via the same tests. We present the general test structures. We focus on *visual attention* tests because we are working in a gaming context, which is a predominantly visual medium. However, the test structures are the same for other modalities, with just the stimuli changed.

Go/No-go Tests (GNG)

Testees respond to stimuli based on a specified behaviour rule (e.g. press the left button when the target stimulus appears). GNGs record testee response times, the number of correct/incorrect responses, and error types. Different error types (i.e. contexts) reflect

inhibition or selective attention. Consider a task where testees must press the space bar when an A is shown on screen, and nothing when an X is shown. Given the sequence A-A-A-A-A-A-X, a testee who presses the space bar on X would be showing a lapse of inhibition. By virtue of their design, GNG also measure processing speed.

GNG rules, response actions, and stimuli vary in context and complexity [252]. We can conceptually categorise GNG as *simple* and *parametric* based on their complexity.

A simple GNG has one behaviour rule and one target. For example, in Konishi et al. [239] a screen shows two square outlines (e.g. left and right squares). In each trial only one square will light up either green or red. When one of the squares is green, the testee must press the key associated to the side of the square as quickly as possible (i.e. if left square is green, press the left key). When there is no green square the testee should not press anything.

A *Parametric GNG* changes the behaviour rules and target set during each level of the test. For example, in Langenecker et al. [252] a screen shows a series of letters one at a time; the testee must do the specified response based on the target letters (Fig. 9.1). The target set and response become increasingly complex as the testee moves up in levels. At Level 1 the target must press the response button whenever the screen shows X, Y, or Z. Level 2 reduces the target set to X and Y; the testee now must only respond when the stimuli alternate (i.e. if you see an x twice, do not press it the second time). Level 3 keeps the behaviour rule the same, but reintroduces Z as a target.

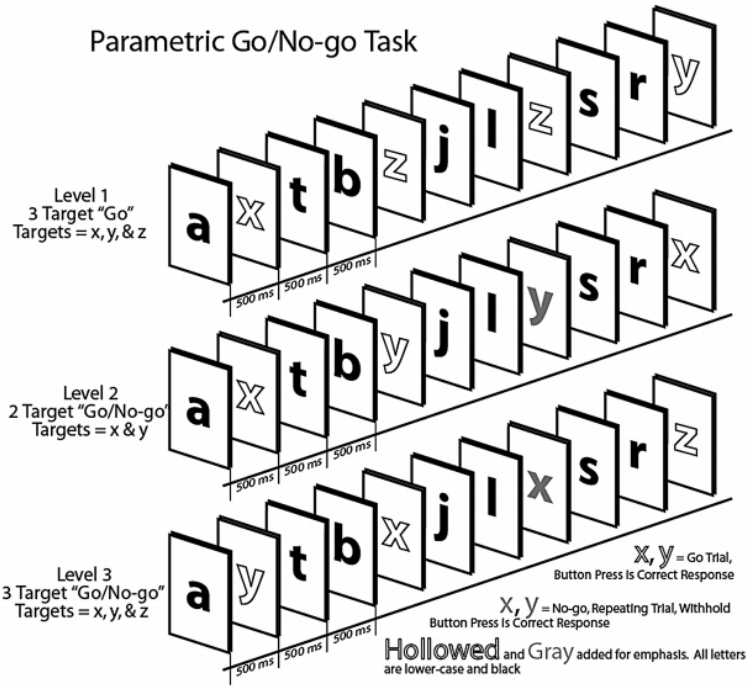


Figure 9.1: Reproduced image of levels in Parametric Go/No-go Task from Langenecker et al. [252].

Continuous Performance Tests (CPT)

CPT are similar to a simple GNG conducted over a longer period of time. Their goal is to force long task-engagement so that testee responses become somewhat automatic, ensuring that errors accurately reflect inhibition and selective attention [259]. CPT record the same metrics, and use the same error type logic to measure inhibition and selective attention. They also measure sustained attention due to their long duration, and processing speed based on response time. We explore three common CPT: the AX-CPT, the Flanker CPT, and the Tests of Variables of Attention.

The AX-CPT [428] displays letter sequences to the testee, who must press one button ("Z" on keyboard) when they see an A-X sequence, and a separate button ("/" on keyboard) for all other stimuli (Fig. 9.2). There are four sequences: one target (A-X) and three distractors (A-*, *-X, *-*; where * represents any letter not A or X). The ratio of targets to distractors varies between AX-CPT instances, but all designs favour target sequences [29]. A single AX-CPT session consists of 4 blocks, each with 10 to 40 trials (individual letters). At its longest, AX-CPT instances can take up to 45 minutes to complete [29].

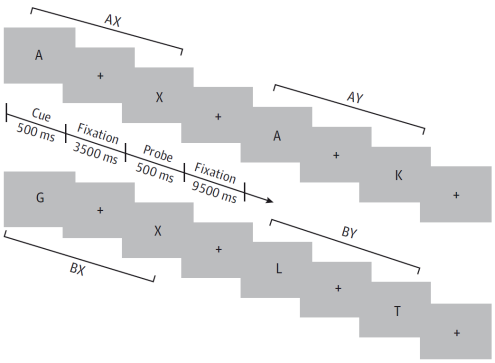


Figure 9.2: Reproduced AX-CPT sequences from Lesh et al. [262].

Flanker CPT (F-CPT) [126] displays five adjacent stimuli (Fig. 9.3), where the third (i.e. middle) image is the target (an arrow) and the other images are distractors (either line segments or arrows)[3]. The testee must respond depending on the direction of the target (e.g. click left mouse button for left arrow, right mouse button for right arrow). F-CPT sessions are 144 trials (one display) long. Trials are either neutral (distractors are line segments), congruent (distractor arrows point in the same wdirection as target), or incongruent (distractor arrows point in different directions). The trial types are randomly distributed, and the experiment is not divided into blocks.

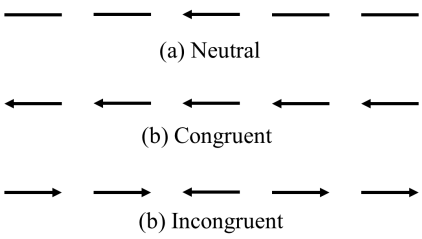


Figure 9.3: F-CPT Trial Types.

Test of Variables of Attention (TOVA) [259] are simple GNG (one target stimulus, one non-target stimulus, testee responds only to target) where the stimuli appear in separate locations (Fig. 9.4). TOVA administers two trial sets: a target-heavy trial (3.5:1 target to non-target [259], or more than 50% target) and a non-target-heavy trial (1:3.5 target to non-target [259], or less than 50% target). The target-heavy set measures inhibition, while the non-target-heavy set measures selective attention. Completing a set takes about 11 minutes for anyone above 10 years-old [259].

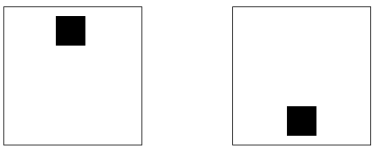


Figure 9.4: TOVA target (a) and non-target (b) stimuli.

[3]The original test used letters, but most versions use arrows because they are culturally neutral [241].

Stroop Colour-Word Test

A two-phase, predominantly analog, test where testees must read-aloud from a written list of colours, either saying the word as written or the font-colour. For example, if given Fig. 9.5b, and asked to read the words, a testee should say "blue, green, red, yellow, gray, black". However, if they are asked to read the colour they should say "yellow, black, gray, green, blue, red". Stimuli are either congruent (word matches the font colour) or incongruent (word does not match font colour), though congruency can also be measured in degrees of similarity (e.g. when the word and its font colour do not match but begin with the same letter like the word BLACK in blue font) [371]. Computerized versions may substitute reading-aloud for pressing a button for congruent stimulus, or pressing a specific button based on the font colour [63, 371]. Like other attention tasks, Stroop tests uses error rates and types to measure selective attention, and inhibition. Computerized versions also collect response times to measure processing speed.

Figure 9.5: Stimuli for Stroop Colour Word Test. Image (a) is an example of congruent stimuli, and (b) is incongruent stimuli.

9.2.4 Object Recognition

We found three types of object recognition tasks (ORT): object detection, object identification, and novel object recognition. The object detection and identification tasks come from Grill-Spector and Kanwisher [171], and the *Visual Object and Space Perception Battery (VOSP)*[514], and use real world objects. Novel object tasks are more varied following the procedure by Richler, Wilmer, and Gauthier [394].

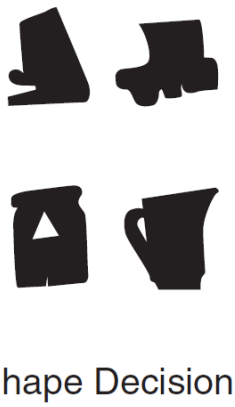


Figure 9.6: Reproduced VOSP example from Carone [82]

Object Detection (OD-ORT)

Testees identify whether the stimulus is a real object or not. They do not have to *name* the object, hence the separation from object identification tasks. For example, Grill-Spector and Kanwisher [171] shows the testee either a target (a real world object) or non-target (a scrambled real object covered in textures or dots) image for a short time (max 167ms), then removes the image and gives testees two seconds to decide and respond via button press. OD-ORT may also present stimuli simultaneously. For example, the *VOSP Object Decision Test* has testees pick out the real object silhouette hidden beside three fakes (Fig. 9.6). OD-ORT record correct and incorrect responses.

Object identification (ID-ORT)

Testees "name" the real world object they are shown (i.e. perceive and recognize the stimulus). ID-ORT "naming" happens implicitly via categorization or explicitly by verbal response. All ID-ORT collect correct and incorrect response data, which is used to calculate accuracy. Some also collect response time as a measure of processing speed.

Categorization Tasks: Testees must quickly and correctly identify a picture to its category. Grill-Spector and Kanwisher [171] presents two main tasks: ***broad categorization***, and ***within-category identification***. Both follow the same structure: testees are shown a picture for up to 167 milliseconds, then the picture is obscured and testees have 2 seconds to identify the stimulus. For **broad categorization** this was identifying the stimuli as either a face, bird, dog, fish, flower, house, car, boat, guitar, trumpet. For **within-category identification** the testees had to identify whether the stimuli was of a specific exemplar of the category (Face: Harrison Ford, Bird: pigeon, Dog: German Shepard, Fish: shark, Flower: rose, House: barn, Car: VW beetle, Boat: sailboat, and Guitar: electric guitar) or not ("other").

Naming the Stimulus: Testees must identify stimulus that are either obscured or in silhouette. VOSP presents four tests: *the Shape Detection Test*, *Incomplete Letters*, *Silhouettes*, and *Progressive Silhouettes*. **Shape Detection** testees must identify if there is an "X" on 20 degraded images (Fig. 9.7a). **Incomplete Letters** has testees identify 20 letters that are degraded by either 70% or 30% (Fig. 9.7b). **Silhouettes** has testees identify 30 increasingly difficult silhouettes of animals or common objects photographed from unusual views (Fig. 9.7c). **Progressive Silhouettes** has testees identify an object (either a handgun, or a trumpet) from an incomplete/degraded silhouette (Fig. 9.7d). Testees who do not identify the object are shown up to ten progressively clearer silhouettes. Progressive silhouettes stands out as it calculates performance as the number of images needed to identify both objects (e.g. if it took 5 images to identify the handgun, and 2 to identify the trumpet the score is 7; max score is 20).

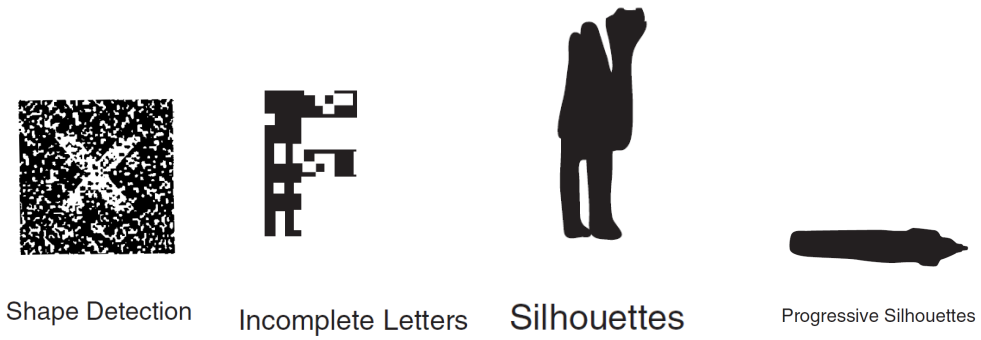
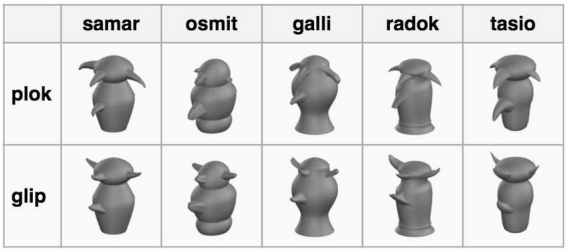


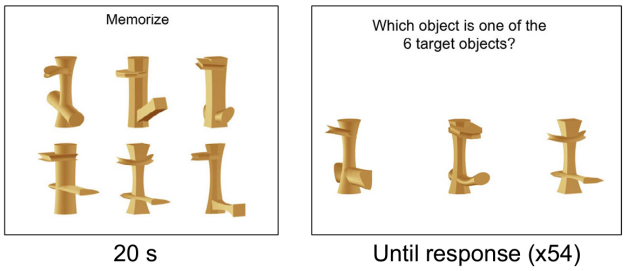
Figure 9.7: Example stimuli from VOSP ID-ORT, reproduced from [82].

Novel object recognition (N-ORT)

The general structure of N-ORT is to show the testee a target image, then after a delay present them with multiple similar options and ask them to select the one they had been shown. The stimuli are generally Greebles [152] (Fig. 9.8a), nonsense figures which each have four parts that systematically vary along shape ("family") or orientation ("gender")[4]. However, there are studies that use rotated and edited images of real world objects (e.g. 385), and physical objects have been used to study OR in mice following a similar process (e.g. 124). N-ORT collect the testee's correct and incorrect responses to calculate accuracy, and response times to understand processing speed.



(a) Greebles organized by gender (plok/glip) and family (samar, osmit, galli, radok, tasio).



(b) NOMT example test using Ziggerins reproduced from Richler, Wilmer, and Gauthier [394].

Figure 9.8: Classical elements of a novel object recognition study.

The quintessential N-ORT is the **Novel Object Memory Test** (NOMT) [394]. Testees start with a learning phase (18 trials) where they are shown three views of an object, and then must pick which object they just saw from three options. The testees are then shown a final view of all novel objects for 20 seconds before the real test begins. For the rest of the test (54 trials) testees must select the target object from the three options (Fig. 9.8b). Rajalingham, Schmidt, and DiCarlo [385] run an N-ORT without Greebles. Testees do not have a learning phase; instead they are shown the test image and are then immediately tested. The whole process can be seen in Fig. 9.9.

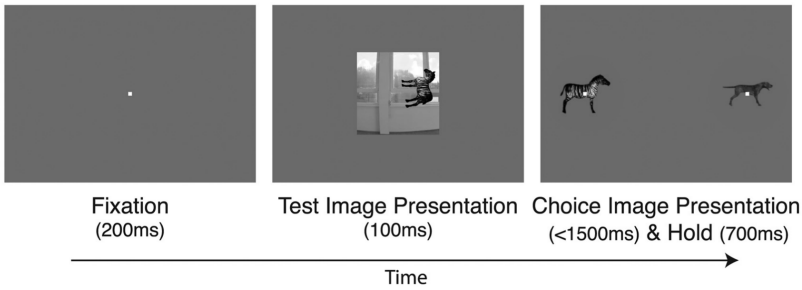


Figure 9.9: Trial structure from Rajalingham, Schmidt, and DiCarlo [385]. Note: their paper states in the body of the text that the fixation is for 500ms, but the image says 200ms.

[4]Not all N-ORT use Greebles as there are other similar stimuli like Ziggerins [530].

9.2.5 Token Change Detection (TCD)

Change detection tasks (CDT) have testees identify changes between an initial scene and second scene. Depending on the implementation, testees may be asked *if* there was a change (yes/no question), or to *identify* the change. Feuerstahler et al. [137] identify three types of CDT: canonical, partial report, and multiple change (Fig. 9.10). All types capture the correct and incorrect responses which they compare to the set size, and number of changes. Response time can also be captured, though CDTs tend to focus on accuracy.

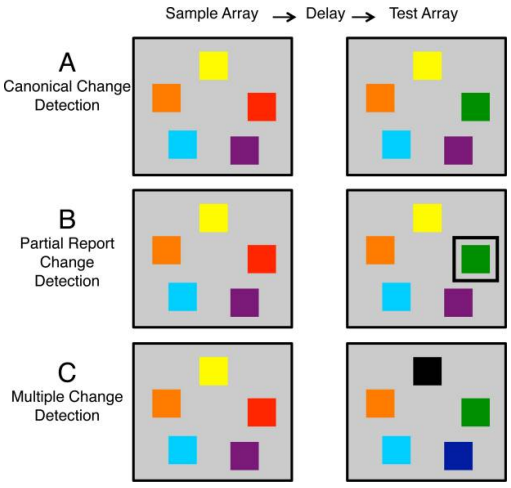


Figure 9.10: Types of CDTs reproduced from Feuerstahler et al. [137].

Canonical CDT (Fig. 9.10a) makes up to one change between scenes and asks testees *if* they notice a change. For example, Vogel, Woodman, and Luck [511] presents a scene of one to twelve coloured squares for 100ms, hides the scene for 900ms and then reveals the second scene for 2000ms. Testees indicate a colour change via button press; about 50% of trials are changes. The total test runs 30 to 45 minutes long.

Partial report CDT (Fig. 9.10b) makes up to one change, and asks testees if they notice a change in a particular element. For example, Maxcey-Richard and Hollingworth [285] show testees a workshop scene one item at a time, with a high or low pitched tone playing alongside each item to indicate the likelihood of this item being the target (high pitch indicating higher likelihood). After presenting 6 items, the test draws a green box around an item in the scene and testees must indicate if the item changed by pressing one of two buttons. Like Canonical CDT, 50% of the trials had changes. Partial report CDTs minimize "decision noise", but could increase errors because of remembering non-cued items and potential location binding (where the features of the object are remembered but its location is lost)[137].

Multiple change CDT (Fig. 9.10c) changes any number of elements between scenes, but keeps the set size (total number of stimuli on screen) the same. Like Canonical CDT, testees only indicate if they think there was a change. Gibson, Wasserman, and Luck [156] use this to compare visual memory of humans and pigeons. Each testee sees an initial scene of 8 objects for 1 second, followed by a delay (i.e. blank screen) for 1 second, and then a second set of objects for 1 second. Human trials change the scene 50% of the time, with an equal likelihood of 1, 2, 4, 6, or 8 objects changing.

9.2.6 Tactile Perception

Tactile perception provides information about the object being held, particularly quality information (e.g. weight, size, number, texture, firmness) and spatial information (e.g. orientation of object, layout of object elements, position relative to us). It is measured through haptic ORT. Most of the these tasks emulate visual ORT, and other perception tasks with the stimuli replaced with something tactile. For example, Grunwald et al. [172] mimic a N-ORT by etching 12 abstract lines (i.e. meaningless shapes) into a metal plate, having testees feel the shapes, wait 10 seconds and then attempt to draw them from memory (Fig. 9.11). They record how long testees spent feeling the shapes, and the accuracy of the drawings (i.e. how many were correctly remembered).

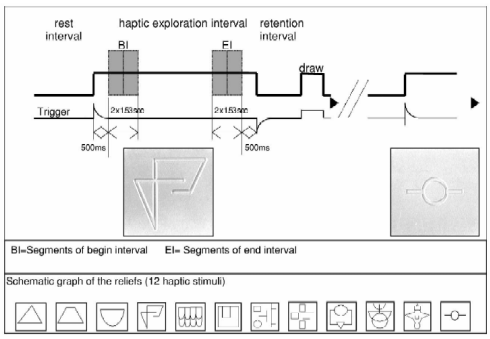


Figure 9.11: Example trial and line drawings, reproduced from Grunwald et al. [172].

9.2.7 Procedural Memory

This ability integrates many other abilities, and so is difficult to isolate. We find three types of tasks that incorporate procedural memory: pursuit rotor tasks, choice reaction time tasks, and mirror tracing task.

Pursuit Rotor Task (PRT)

PRT is a motor coordination task where a target circle moves along a circle track at a fixed speed and the testee must keep their cursor on the dot [313]. There are two sets of four trials each; the first set asks testees to use their dominant hand, while the second asks for the non-dominant. Each trial is about 15 seconds (though implementations may vary), bringing the whole test to be about two minutes long. The test records error magnitudes (how far off target) and times (when errors occurred). Procedural memory is measured through the errors; the task involves predicting the movement of the target based on recalling how it moves and matching your motor movement to it.

Choice Reaction Time Tasks (CRTT)

CRTT are a family of tests similar to GNG, where testees perform a context-based response action to a stimuli (e.g. choosing which button to press in responses to a target image). CRTT vary along the number of choices presented, types of responses, and target properties (e.g. placement, type, modality, etc). Like GNG, it measures abilities through error types and contexts. A common CRTT is the **Serial Reaction Time Task (SRTT)**, where testees push one of four buttons in response to the stimulus appearing in a specific location on the display. Target locations may be in a row (e.g. 489) or in the four cardinal directions (e.g. 426). Responses can also be tied to the target type (e.g. 233). SRTT are often divided into

multiple trial sets; one set being randomized, with the others following repeating patterns like GNG.

Mirror Tracing Task (MTT)

Testees trace an on-screen image with their mouse, while the controls are reversed (i.e. moving the mouse left moves the cursor right). There are three trials: two practice trials where testees trace a line and an L respectively, and a final trial where testees must trace a star shape. Measures are often in the form of time taken to trace the shape, and errors (deviations from the tracing line). Implementations vary shapes to trace and whether the task is timed. For example, Renna et al. [389] measures persistence by limiting the trials at 60 seconds for the practice trials, and 7 minutes for the star.

9.3 Summarizing Player Ability Measurements

From this survey, we have a better understanding of how to measure each ability. However, many of these measures are part of proprietary analog batteries. They also require specialized equipment, or were generally only feasible in a strictly controlled lab setting.

We realise that we need to build our own test battery to gather all of these measures in one place. This poses a challenge as some abilities may be quite difficult to measure without specialized equipment (e.g. Steadiness test), or very long testing sessions (e.g. continuous performance tasks). Luckily we also find that many of the task designs for cognitive abilities overlap, and often collect the same metrics. With this understanding we move into Ch. 10 where we choose which measures to implement, and explore how to do that.

Take home points

From this chapter we learned the following meta-lessons:

- We focus on abilities specific to button mashing to make their exploration more feasible and likely to produce a complete model.
- We identify 24 unique test types over 6 abilities (Table 9.1).
- Motor measures often require specialized equipment that may be difficult to replicate digitally.
- Cognitive measures significantly overlap in task design and metrics.

Chapter 10

Designing the Mini-Game Battery

We want to create a test battery which provides the information we need to construct a player profile. We decide to create our own versions of the ability measures from Ch. 9. This allows us complete control and flexibility over the task variables and outputs.

As we aim to model players in a gaming context, it seems appropriate to implement this test battery as a *game-based assessment* (GBA). GBAs are games which build their core mechanics and challenges around the measurement [250], such that players only perceive the game and not the assessment. There has been some evidence that GBAs engage testees more than standard digitized assessments [250, 268] and decrease test anxiety [250, 284]. Using GBAs would encourage players to tap into their existing gameplaying schema, potentially giving us insight into their at-home game playing behaviour and a more realistic understanding of their abilities when playing games. For our test battery, we can use our knowledge of the measurement tasks from Ch. 9 to design a series of mini-games targeting specific abilities.

This chapter covers the design of our test battery mini-games. We start by identifying project constraints, and setting criteria for the ability measures we can implement. We use these to prune our potential tests (Table 9.1) into a shortlist that could be reasonably translated into games. We then explain our role in the design of the test battery. **Sasha is "*project supervisor*" on the design and implementation of the mini-games and battery framework. The test battery implementation is the work of several undergraduate level students for their Capstone project, and M.Eng student, Vansh Pahuja.** We then provide a high-level description of the mini-games, summarizing their target ability, base test, and metrics. The design details and inspirations are in App. A.2 to keep the body of this thesis concise.

10.1 Constraints and Criteria for Test Selection

Our survey outlines 24 tests (Table 9.1), many of which require specialized equipment, or very long sessions. Our test battery aims to measure abilities through a series of mini-games playable in a single session. It is unreasonable to subject a player to *all* of these tests (especially multiple ones that take thirty minutes or more!). To this end we outline criteria to refine the test list, and identify constraints on the battery implementation which further reduce our options.

10.1.1 Constraints

The design and development of the test battery framework occurs during the height of the Covid-19 Pandemic. Campus closures and lockdowns impact our ability to use lab equipment for developing and validating the battery. Since it is unclear how long these physical constraints will be in place, we impose the following design constraints on the project:

- the tests will be run on computers since we cannot run something on a console; and,
- the interface will be limited to keyboards because we cannot guarantee that participants have an external device (e.g. controller) or even an external mouse (using a laptop trackpad will skew the results between participants because of the method of interaction).

Essentially we can only construct tasks that are digital and use discrete, button-based input. This means that options like the pursuit rotor and mirror tracing tasks are not possible. Trying to replace fluid motor movements that are achieved either by joystick or freehand with combination button movements changes the complexity of the task and is less accurate of a measure.

A Quick Aside...

COVID-19 campus closures ended after the development of the battery framework and initial mini-game implementations. Even though these constraints may not seem necessary now, they significantly affected this project's scope and how we could validate the battery.

10.1.2 Criteria

We consider the criteria when selecting tests:

Cr.1 Tests should be directly applicable to the gameplay context we're testing;
Cr.2 Tests should be relatively short in order to not make the overall battery excessively long;
Cr.3 The final selection of tests should not have unnecessary redundancy (i.e. we do not need several tests for the same ability when one test with multiple levels of difficulty could suffice); and
Cr.4 Tests should be easily adaptable into games.

These criteria support the gaming context of our assessments, and keep the overall length of the battery short.

Gaming context: Cr. 1 and 4 aim to strengthen the "game feel" of the battery. Cr. 1 says it is important for the task structure and goals are meaningful in a game context. For example, the *Object Detection* task relies on "real" and "fake" objects; however, players frequently encounter "fake" objects (like impossible weapons, and items) in games that are

meaningfully real in the context of their play. Therefore it may not be a useful structure for our work. Cr. 4 supports this by asking us if this structure exists in commercial games. If a task seems similar to existing mini-games it means that players already have schemas for engaging with it and we may get more reliable measures. For example, the mail sorting mini-game in Legend of Zelda: The Wind Waker [327] and Sort of Fun in Super Mario Party [324] both use the structure of *Categorization tasks* (Fig. 10.1). These criteria together make us more inclined towards categorization tasks over detection tasks for measuring object recognition.

(a) Mail sorting in Legend of Zelda: The Wind Waker [327]

(b) Sort of Fun in Super Mario Party[324]

Figure 10.1: Examples of commercial gameplay that mirrors the structure and goals of categorization tasks.

Time: Cr. 2 and 3 are useful for keeping the battery length short. Given a set of tests for an ability, these criteria lead to picking one short version of tests. For example, selective attention and inhibition have six potential tests: two Go/No-Go (Simple GNG, Parametric GNG), three continuous performance tasks (AX-CPT, Flanker, TOVA), and the Stroop test. Continuous performance tasks are long (often over 20 minutes) so we can eliminate them as options (Cr. 2). We now have three to pick from (Simple GNG, Parametric GNG, and Stroop). Considering Cr. 3, we lean towards the Parametric GNG as its structure allows for more variation in difficulty type and level through changing target stimuli and response behaviours. We can also apply these to reduce overall games needed. Consider procedural memory, a parametric GNG is incredibly similar to a choice reaction task (CRT). A game designed as a parametric GNG could easily output information that is also applicable to procedural memory, reducing the overall length of the battery.

10.2 Selection of Tests for Ability Battery

We use our criteria to arrive at a refined list of tests (Table 10.1). These tests serve as a guide for the game design, and a starting point for the teams developing the mini-games. Not every test will be implemented, as there is still some redundancy in the list.

[1]Errors are treated separately from incorrect responses because it covers all types of errors.

Ability	Tests	Metric
Finger pressing	Finger tapping test	No(P)/10s
Selective Attention & Inhibition	Simple GNG	RT_{ms}, No(C), No(I), ErrT
	Parametric GNG	RT_{ms}, No(C), No(I), ErrT
	Stroop Colour-Word Test	RT_{ms}, No(C), No(I), ErrT
Object recognition	Object Categorization (broad)	No(C), No(I), Acc
	Object Identification	No(C), RT_{ms}
	Shape Detection	No(C)
	Incomplete Letters	No(C)
	Silhouettes	No(C)
	Progressive Silhouettes	No(C)
Token change detection	Canonical CDT	No(C)
	Partial Report CDT	No(C)
	Multiple Change CDT	No(C)
Procedural memory	SRTT	RT_{ms}, No(Err), ErrT
Tests: Go/No-go task (GNG), Change detection test (CDT), Serial reaction time (SRTT)		
Metrics: Number of X (No(X)), Presses (P), Correct responses (C), Incorrect responses (I), Accuracy (Acc), Errors (Err)[1], Types of Errors (ErrT), Response time in X (RT_X), Seconds (s), Milliseconds (ms)		

Table 10.1: Viable ability tests on which to model mini-game ability battery tests.

Explaining Cuts: All tests for wrist pointing and tactile perception, and two tests from procedural memory (pursuit rotor and mirror tracing) are eliminated because they require specialized equipment we could not access (i.e. Constraint reason). All continuous performance tasks (CPT) are eliminated due to their lengths (i.e. Criteria reason - Time). Object Detection and Object Decision tasks are eliminated because they focus on *if* something is an object, which is less important in a gaming context than identifying *what* an object is (i.e. Criteria reason - Gaming context). We eliminate novel object recognition tasks because, when considering their structure in a gaming context where all items may be novel, they overlap considerably with identification tasks (i.e. Criteria reason - Gaming context & Redundancy).

10.3 Design of Tests for Ability Battery

With a list of potential tests we could start designing the mini-games and building the battery framework. Sasha supervises two undergraduate capstone groups in the production of the battery framework. She also supervises one Masters of Engineering student in the creation of several mini-games. Sasha's contributions to these projects are:

- project scoping and direction,
- defining requirements and constraints,
- providing initial inspirations leading to game designs (particularly for Masters student),

- making sure mini-game design reasonably follow ability measures and providing design advice to this end,
- game design testing,
- empirical validation of games (see Ch. 11 for details)

The result is seven unique games, five of which are relevant to our research: Digger, Stage, Cake, Recipe, and Looking. The other two (Feeder, and Rockstar) target non-button mashing related abilities, and so do not fit our work. The five relevant mini-games emulate various commercial mini-games modified to implement a specific ability test for one button mashing ability. A sample of commercial gameplay that aligns with ability tests is covered in App. A.1. Table 10.2 summarizes the gameplay, target ability, test inspiration, and a Mario Party mini-game example for each implemented game[2]. Each game outputs its all interaction events, so we know the button pressed, the event time, and trial. From these we can calculate the appropriate metrics for each ability based on the test base. Details about the game designs and parameters, along with links to the GitHubs can be found in App. A.2.

10.4 Summarizing Mini-game Battery Design

In refining the list of ability tests from 24 to 14 we were able to get a better idea of how to design games in our test battery. We pull on our game knowledge to connect existing mini-games to ability measures, thus creating inspirations for the games we could make. Our work here is the foundation for two capstone projects and a Masters project, the results of which are an implementation of our test battery and seven unique mini-games. Future work can expand on this battery to incorporate different types of tests now that we are no longer limited by Covid-19 restrictions.

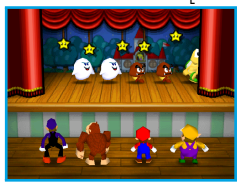
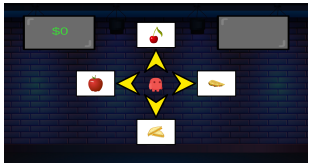
For our thesis, we now have to validate the test battery. We aim to show that it can reasonably measure player abilities, and therefore can be used to create player profiles.

Take home points

From this chapter we learned the following meta-lessons:

- Our test battery is a game-based assessment to get a more realistic view of ability use in gaming contexts.
- We had to refine our list of ability tests to ones that met our time (i.e. playable in a single session) and gaming context goals, and our COVID-19 related constraints.
- We supervise the design and implementation of the testing battery framework, and 7 unique ability-testing mini-games (the 5 relevant to thesis are presented in App. A.2).

[2]Each game was inspired by *many* mini-games and other commercial gameplay moments. This just serves to illustrate the various ways the games are similar.

Game	Description	Ability	Base Test	Mario Party Examples
	Digger: Mash the button to dig for treasure.	FP	FTT	Will Flower [197] 
	Stage: Watch the actors walk on and off the stage; decide if there were any changes to the actors when they come back.	TCD	C-CDT	Curtain Call [195] 
	Looking: Select the target item from the set, or press X if it is not there.	SelAtt & Inhib	GNG	Looking for Love [324] 
	Cake: Sort the food on the conveyor belt into the correct bin.	OR	OCT	Sort of Fun [324] 
	Recipe: Sort the correct pair of candies into the shipping box, and place all others in the recycling.	OR, SelAtt & Inhib	ID-ORT, GNG	X-Ray Payday [323] 

Legend: Finger Pressing (FP), Token change detection (TCD), Selective Attention (SelAtt), Inhibition (Inhib), Object recognition (OR), Finger Tapping Test (FTT), Canonical Change Detection Test (C-CDT), Go/No-Go Test (GNG), Object categorization task (OCT), Object Identification task (ID-ORT)

Table 10.2: Summary of games implemented for battery.

Chapter 11

Player Profiling in Action

We aim to show that the Mini-Game Ability Battery (MGBatt) is a reasonably ***valid*** (i.e. accurate) and ***reliable*** (i.e. precise) measure of player abilities. We run a validation-by-correlation study to show sufficient *convergent validity* (i.e. evidence of measures capturing the same constructs [81]) between MGBatt and the Psychology Experiment Building Language (PEBL) [310] battery tests. PEBL is widely used [311], and most of its test implementations are validated (e.g 227, 374), making it sufficient for our purposes. We also test MGBatt for *test-retest reliability* (i.e. consistent agreement of test-retest scores for the same individual measured at different points [8]) as part of this study.

We start by explaining the rationale for our study, and formalizing our validation and reliability conditions (hypotheses). We then outline the study design, procedures, and other relevant information. We present the results of our correlations, with details in App. B. We discuss our (mixed) results, focusing on potential explanations, study limitations, and directions for follow-up studies. We conclude the study by deciding which mini-games will be used to create player profiles.

A Quick Aside...

It is important to recall that the *actual* goal of this thesis is showing that if we compare competency profiles to player profiles we can predict mechanical achievability (i.e. the jutsu concept). Validating MGBatt as sufficiently measuring player abilities is just a stepping stone to validating jutsus.

11.1 Why a correlational study?

We use a *validation-by-correlation paradigm* to test convergent validity between MGBatt and PEBL, and MGBatt's test-retest reliability. Validation-by-correlation is common across domains for showing convergent validity between different measurements (e.g 158), and is the basis of factor analysis which is used to validate many tools (e.g. 4, 232). We aim to show that:

1. MGBatt scores reasonably correlate with PEBL scores (*validity by correlation*); and,

2. MGBatt scores reasonably correlate between tests (*test-retest reliability*).

11.1.1 Measures of Correlation

In order to check for validity and reliability we need to establish how we will measure correlation.

Validity: We practically consider validity to mean the participant with the best performance (i.e. score) on the PEBL measure ought to have the best performance on the associated MGBatt game. The overlap between cognitive abilities, individual differences between participants, and human variability between tests, can all impact the strength of a linear correlation between measures. Given these factors, it seems more reasonable to look for rank correlation between scores (as is common in developmental psychology [182]). To this end, we use ***Spearman's Rank Correlation (ρ)* as our validity coefficient**. Unlike *Pearson's* r[1], Spearman's ρ makes no assumptions about data normality, and does not try to fit the data to a linear model, just a monotonic one [407]. We choose to also report r to get an idea of whether any linear correlation may exist between the mini-games and existing measures.

Reliability: We practically consider reliability to mean an individual's MGBatt scores reasonably correlate, and the rank order of the participants is consistent. While intuitively we would use Pearson's r, there is documented concern about its ability to detect agreement between measures [8, 240, 539]. *Intraclass correlation* (ICC), which represents the statistical similarity of multiple ratings within individuals [8], is a common alternative but is susceptible to bias when sample populations have little to no variability [182]. **Considering our view of reliability it seems appropriate to capture both Pearson's r and ICC as our reliability coefficients**. We calculate r as normal to gauge general correlation between the test and retest as an adhoc "consistency" check. ICC calculations differ depending on whether it is capturing consistency or absolute agreement [240]. Following guidelines from Koo and Li [240], we use IBM SPSS Statistics (SPSS version 29) to calculate ICC using a two-way mixed effects model for absolute agreement with a 95% confidence interval. By comparing these two values we will judge whether reliability is sufficiently met for our games.

11.1.2 What is "reasonable correlation"?

The bounds of "sufficient" or "reasonable" correlation vary between contexts [81]. Fraenkel, Wallen, and Hyun [142, p. 334] state that generally validity coefficients should be at least 0.5, and reliability coefficients should be at least 0.7, with higher being better for both. Hedge, Powell, and Sumner [182] note that "good" reliability reports often range from 0.6 to 0.9, and that researchers generally call any coefficient reported "adequate" or "satisfactory". Koo and Li [240] rule of thumb is that, given at least 30 heterogeneous samples and 3 raters (i.e. 3 separate measures), ICC values less than 0.5 indicate poor reliability, 0.5 to 0.75 indicate moderate reliability, 0.75 to 0.9 indicate good reliability, and greater than 0.9

[1] Pearson's r requires the data is reasonably parametric and the relationship is linear [407].

indicates excellent reliability. We use these base values (r and ρ: 0.7, ICC: 0.5) as target thresholds, but rely on our context to define the lower bounds of "reasonable".

MGBatt is for player profiling in a game design context, so we do not hold it to the same levels of scrutiny as clinical measures. We consider our data set to be sufficient for analysis with a minimum 30 data points per test[2]. We set our analysis confidence level at $\alpha = 0.05$[3]. Given this context, we deem the lower bound for "reasonable" Pearson and Spearman correlations are their critical significance values: Pearson $r > 0.349$ [513] and Spearman $\rho > 0.306$. We adhere to the ICC coefficient rule of thumb [240] to gauge appropriateness of ICC values[4]. We summarize our goals into specific hypotheses statements:

Hypotheses:

H1 The MGBatt is valid if the participants' MGBatt scores positively correlate to the associated PEBL scores at $\rho \geq 0.7$.

H2 The MGBatt is reliable if participants' MGBatt scores between sessions positively correlate at $r \geq 0.7$ and ICC ≥ 0.5.

We consider correlations above critical significance values ($\rho > 0.306$ and $r > 0.349$) to be significant (and thus worth further investigation).

11.2 Study Design

We use a within-subject design to generate two paired datasets (MGBatt-PEBL scores, and MGBatt-MGBatt scores).

11.2.1 Apparatus

Lab setup: The experiment is run in the G-ScalE lab on McMaster University's main campus (Information Technology Building, Room 128). The room is setup to have two participants run the experiment at the same time (Fig. 11.1). The two testing stations are oriented such that participants cannot see each other. The lab also has a rest area set up to emulate a living room, with a set of couches, a coffee table and a television.

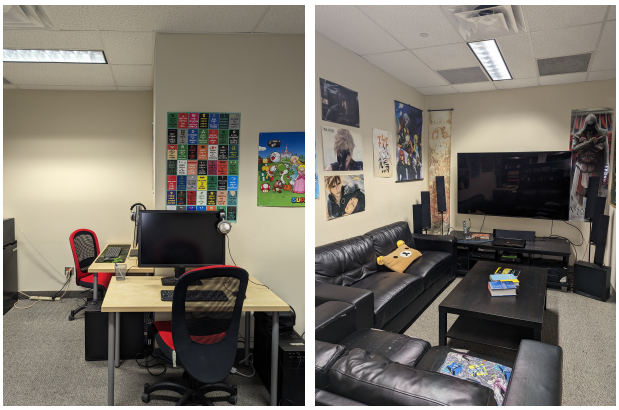


Figure 11.1: Lab setup.

[2] 30 samples is the point where data will begin to display normal distribution qualities (i.e. Central Limit Theorem) and so is a rule of thumb minimum for many research fields.

[3] Meaning that p-values ≤ 0.05 are considered significant.

[4] While p-values are generated for ICC values, they are considered irrelevant [8]. Instead confidence intervals are more important in reporting [8], and they are fixed at 95% in our calculation.

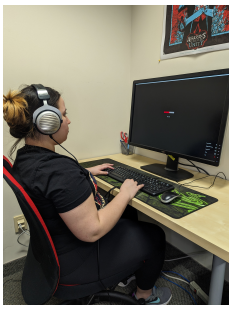


Figure 11.2: Example of participant at a testing station.

Hardware: Each testing station is equipped with a Windows 10 desktop (i5-4670k 3.40 GHz processor, 16 Gb RAM, Nvidia Geforce GTX 780Ti graphics card), 30-inch monitor, a full size wired USB keyboard (K120-TAA Logitech), wired USB mouse (B100-TAA Logitech), and a pair of wired over-the-ear headphones (Beyer Dynamic DT 990). Participants are positioned 26-inches away from the monitor to maintain ergonomics. Fig. 11.2 visualizes this setup. In keeping with Covid-19 safety guidelines, the testing stations are wiped down with disinfecting wipes between participant sessions. Headphone ear muffs are also covered with disposable caps to ensure sanitary use.

MGBatt: The mini-games descriptions are in Ch. 10 and App. A.2. Participants play three sets of five games. Each set of games is tuned easy, medium, or hard as per Table 11.1. The increasing difficulty taxes each ability more, and captures more robust measures. Participants always start with the easy set, then the medium, then hard. We randomize the game order inside the set. Participant output is saved in JSON data files.

Game	Variables	Easy	Medium	Hard
Digger	DigAmount	35	60	80
Looking	AverageUpdateFrequency	18	6	3
Recipe	AverageUpdateFrequency	18	6	3
Cake	AverageDispenseFrequency	2	1.5	1
	FoodVelocity	2.5	3	3.5
Stage	DiffLevel	1	2	3

Table 11.1: Configuration settings for Mini-games at different difficulty levels.

PEBL [310]: We select six PEBL tests, implemented as per the PEBL Manual [309] and the PEBL Wiki (Table 11.2). We choose these tests because of their focus on one of our target abilities, either by being a digital implementation of an analog measure or via task design. To reduce overall PEBL length to match MGBatt run time, we adjust the number of trials per segment in any task over 200 seconds — this is a cut-off of just over 3 minutes per test. We randomize the task order for each participant.

11.2.2 Procedure

Pre-session: Potential participants complete our Pre-Study Survey on Google Forms (App. B.2) which collects their consent, demographic information and gaming history. In this survey participants are informed this is an ***unpaid study***. We screen participants for eligibility on age (18-64 years old) and gaming history (play games at least once a week). We contact eligible participants to schedule them for an in-lab validation session.

[5]While this test's response time is unlimited, we assume participants will respond within 1 second based on our pilot testing.

PEBL Test	Abilities used	Description	Time
Speed tapping test	Fine motor skills (**Finger pressing**)	Tap a key as quickly as possible.	80s
Four choice response time	Reaction time, **Selective Attention**, Visual Processing (Object recognition and visuospatial sketchpad)	Respond to a plus sign that appears in one of the four corners of the screen	125s
Flanker Task	**Selective attention**, Reaction time, **Inhibition**	Make direction response with distraction	387s
Object Judgment (Invariant)	Visual Processing (**Object recognition**, and visuospatial sketchpad)	Determine whether two polygons are identical while manipulating shape, orientation, size.	150s[5]
Object Judgment (Absolute)	Visual Processing (**Object recognition** and visuo-spatial sketchpad)	Participants are shown a polygon, then after a delay they are shown two more and have to select which of the two new polygons match the original	188s
LuckVogel	Change detection (**token change detection**), Working Memory capacity	When shown two scenes with a short delay in between, select whether the second scene is the same or different from the first.	480s

Table 11.2: List of PEBL tests used for our correlational study. Information in this table comes from The PEBL Manual V2.0 [309] and the test implementation in PEBL 2.0 [310]. **Bolded text** indicates the main ability we are using the test to study. Time is measured in seconds (s).

Validation Session: We greet participants and lead them to their testing stations. We assign participants a letter-group to determine the order of their measurements (Group A: PEBL then MGBatt, Group B: MGBatt then PEBL). After completing their first measurement, participants are given a 5-minute break in the waiting area. Participants then return to their testing station to complete the second measure. Once the session is complete, we thank participants for their time and schedule the reliability session for one week in the future.

Reliability Session: We greet participants and set them up at the same testing station with the same configuration file from their validation session. We treat participants as a single group since they are only completing the MGBatt. Once the session is complete, we thank participants again and they are free to leave.

11.2.3 Participants

31 participants (20 male; 11 female) complete our validation session; 16 are in Group A, 15 in Group B. Three (2 male; 1 female) could not return for their reliability session. Participants ages range from 18 to 30, with a mean of 23.0 years (σ 3.9). 5 self-identify as having some form of disability that may impact their performance in tasks requiring attentional control and executive functioning. 87% of participants report playing games for more than 10 years.

Most participants play frequently during the week (42% 2-4 times a week; 48% 5 or more times a week) for roughly an hour each session (48% 1-3 hours; 35% 0.5 - 1 hour)[6]. They commonly play on computer (87%) and smartphone/tablet (65%), and report preferences towards Action[7] (65%) and Roleplaying Games[8] (48%) — though Action-Adventure and Strategy[9] are close (39% each).

11.3 Study Results

We now present the results of our study. We start by explaining our analysis method and then present the validity and reliability results. For conciseness, we present the results in tabular form. We then discuss only the significant results and general reasons why we believe the results look this way. Full data analysis and discussion can be found in App. B.4.

11.3.1 Analysis Method

We collect data electronically via PEBL and MGBatt. Our analysis process involves:

1. Initial preparation and cleaning,
2. Data processing,
3. Statistical analysis

Initial Preparation and Cleaning

Process raw data: We use custom Python scripts to process MGBatt's raw event data into accuracy and reaction time values for each trial. This matches PEBL's data format, to allow for easy analysis. The only exception is Digger and Tapping which report individual button press information; we report these in press rates (presses/second).

Organize data sets: We pair MGBatt and PEBL data as per Table 11.3, such that we have a validation dataset with every participants' initial MGBatt and PEBL scores for all games/tasks. We then create a second data set for reliability data, such that each participant's initial MGBatt score is paired with their retest scores. We clean these sets by removing pairs with missing data.

Data Processing

Selecting performance measures: We identify appropriate performance metrics for each ability (Table 11.3). We select these based on task design. For example, in a game where like Looking both speed (i.e. reaction time) and accuracy (i.e. correct responses) are important to evaluating the ability, compared to a game like Cake where the speed is fixed so only accuracy matters.

[6]Inspite of this only 61% of participants self-identify as gamers (70% of males; 45% of females).
[7]Code includes shooters, platformers, fighting games, party games, and pure rogue-likes.
[8]Code includes MMOs, RPGs, JRPGs, Rogue-likes/Rogue-lites with strong RPG elements.
[9]Code includes RTS, MOBAs, Tower Defense, and Base-building games.

Game	PEBL Test	Ability Targeted	Performance Measure
Digger	Tapping	Finger tapping/ Finger pressing	Press rate (press/ second)
Cake	Object Judgment (Invariant) Object Judgment (Absolute)	Object recognition	Accuracy
Recipe	Flanker Four Choice	Inhibition Selective attention	Rate Correct Score (correct responses/ second)
Looking	Flanker Four Choice	Inhibition Selective attention	
Stage	Luck Vogel	Token Change Detection	

Table 11.3: Tested Mini-game and PEBL Pairings and their measurements.

A Quick Aside...

We originally analysed just accuracy measures for all cognitive abilities. However, accuracy measures are only reliable for showing individual differences when participants make sufficient errors, and the speed-accuracy trade-off is controlled for by fixing reaction times or making them irrelevant [118]. The game/task designs make accuracy scores not useful for games except for Battery's Cake. We ran the accuracy analyses anyway so details can be found in App B.3.

Transforming Data: We integrate accuracy and reaction time data into a *rate correct score* (RCS) [529]:

$$\begin{aligned} \text{RCS} &= \frac{\text{Accuracy}}{\overline{\text{RT}}} \\ &= \frac{\frac{\text{Number correct responses}}{\sum \text{Trials}}}{\frac{\sum \text{RT}}{\sum \text{Trials}}} \\ &= \frac{\text{Number correct responses}}{\sum \text{RT}} \end{aligned}$$

This represents the number of correct responses per second in each game, and can meaningfully capture differences in participant choices for speed-accuracy trade-off. RCS has been shown to be a valid way to integrate reaction time and accuracy data (e.g. 495).

Statistical Analysis

Custom Python scripts using Pandas, SciPy, Statsmodels, and Seaborn were used to perform descriptive and correlation analysis on the cleaned data and generate graphs. SPSS version 29 was used to run ICC calculations. Given our participants could have individual differences

which make significant deviation plausible and reasonable, we decide to handle outliers on a per case basis and leave them in the analysis. Ad-hoc graphs and analysis were done in Excel and SPSS version 29 when necessary. All custom Python files are available on Sasha's GitHub for review.

11.3.2 Validity Results

All datasets meet our minimum sampling requirements. Due to a PEBL measurement error in Tapping, and Battery measurement error in Recipe those datasets have only 30 available data points. Table 11.4 reports the correlation results using all 30 data points. Detailed discussion about the data can be found in App. B.4.

Pair		Validity (rounded to 2 decimals)		
Game	**PEBL**	**No. Obs.**	**Spearman ρ**	**Pearson r**
Digger	Tapping	30	0.77^{***}	0.74^{***}
Looking	Four Choice	31	0.38^{*}	0.37^{*}
Looking	Flanker	31	0.21	0.42^{*}
Cake	Object Judgment (Invariant)	31	0.36^{*}	−0.02
Cake	Object Judgment (Absolute)	31	−0.01	−0.09
Recipe	Four Choice	30	0.40^{*}	0.44^{*}
Recipe	Flanker	30	0.34	0.30
Stage	Luck Vogel	31	0.44^{*}	0.56^{**}

Significance: not significant ( ), $p < 0.05(^{*})$, $p < 0.005(^{**})$, $p < 0.0005(^{***})$

Table 11.4: Summarized validity results from correlation analysis of minigames and PEBL tests. Rows highlighted in green support our validation hypothesis, rows highlighted in yellow show positive correlation and require further discussion.

Recall we focus on the Spearman correlations, with evidence of validity at $\rho \geq 0.7$ and significant correlations worth further investigation at $\rho > 0.306$ (assuming $p \leq 0.05$). Digger and Tapping strongly correlate ($\rho = 0.77$, $p < 0.0005$), suggesting it is a valid measure of Finger Pressing. The correlations for Looking-Four Choice ($\rho = 0.38$, $p < 0.05$), Recipe-Four Choice ($\rho = 0.40$, $p < 0.05$), Cake-Object Judgment Invariant ($\rho = 0.36$, $p < 0.05$), and Stage-Luck Vogel ($\rho = 0.44$, $p < 0.05$) cross our significance threshold. We interpret this to mean Looking, Cake, Recipe and Stage may have a stronger relationship than we could detect in this small-scale study, and are worth investigating further. Flanker did not correlate with either Looking ($\rho = 0.21$, p = n.s) or Recipe ($\rho = 0.34$, p = n.s). Cake and Object Judgment (Absolute) did not correlate (ρ = -0.01, p = n.s).

11.3.3 Reliability

We do not meet our minimum sampling requirements for reliability correlation. As well, during data cleaning participants with measurement errors are removed, causing some pairs

to dip below the expected 28 observations. Therefore we **cannot draw any conclusions about the reliability of the mini-games or how they compare to our hypothesis**.

However, from basic looks at Pearson r, Spearman ρ, and ICC[10] (summarized in Table 11.5), Digger ($r = 0.83$, $p < 0.0005$; ICC = 0.79, $p < 0.005$) shows promise of being reliable ($r \geq 0.7, ICC \geq 0.5$). Looking meets the ICC threshold (ICC = 0.58, $p < 0.005$), but not the Pearson threshold ($r = 0.66$, $p < 0.0005$). Cake almost meets both the Pearson ($r = 0.68$, $p < 0.0005$) and ICC (ICC = 0.49, $p < 0.005$) thresholds. Stage and Recipe do not show reliability.

Game	**No. Obs**	**Pearson** r	**Spearman** p	**ICC**
Digger	27	0.83^{***}	0.81^{***}	0.79^{**}
Looking	26	0.66^{***}	0.60^{**}	0.58^{**}
Recipe	27	0.60^{**}	0.68^{***}	0.39^{**}
Cake	28	0.68^{***}	0.49^{*}	0.49^{**}
Stage	28	0.43^{**}	0.32	0.41^{*}

Significance: not significant (), p <0.05 (*), p <0.005 (**), p <0.0005 (***)

Table 11.5: Summarized reliability results from correlation analysis of mini-games vs mini-games measured one week apart.

11.4 Discussing the Validity Data

Since we cannot draw any conclusions about reliability due to the insufficient sample size, we focus this discussion on the validity results.

The most obvious reason for our low-correlations, and therefore mixed results, is the limited sample size leading to low power in the study. This is an unpaid study that required participants to come into the lab for two-sessions, one-week apart. The lack of compensation limits the number of participants, as this commitment is unappealing without incentives. We tried to accommodate for this by setting our sample minimums, but overall a larger sample size could improve correlation strength and provide clearer results.

11.4.1 What about data clustering?

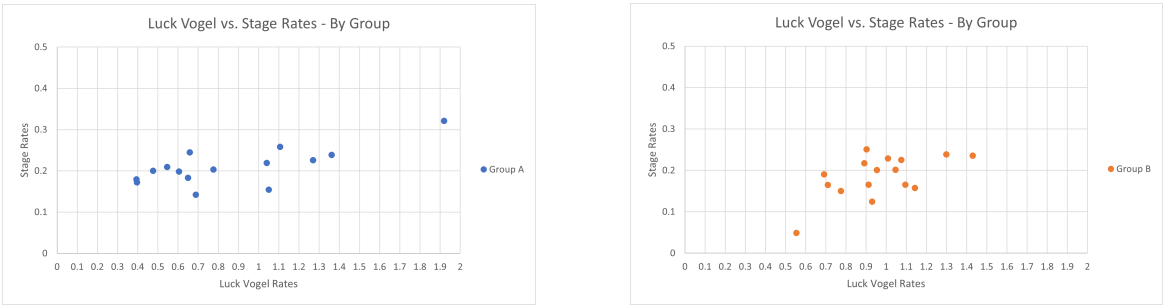
If we decide to take this low sample size data at face value, the low correlations could be due to the significant data clustering we see in the scatterplots (found in App. B.4). We think some potential reasons for this are: a fairly homogeneous sample clustered around peak performance, the potential for priming based on task design overlap, poorly tuned task difficulty, and ability isolation in task design.

Homogeneous Sample: Due to recruitment criteria and location we have a fairly homogeneous sample (relatively able-bodied, university educated, between 18 and 30, significant

[10]Recall ICC is calculated as an absolute agreement coefficient using SPSS via a Two-Way Mixed Effects, Single rater paradigm. This is in accordance with guidelines for test-retest reliability [240]

gamers). This is compounded by the participants' general preference towards action-games (a genre requiring quick reactions, decision making, and other abilities that overlap with those we measure). There is evidence their gaming history could lead to higher performance in assessments of cognitive abilities [392]. Similar abilities, leads to similar measurements, leads to clustering.

Priming: Since MGBatt and PEBL are measuring the same abilities, participants may be primed by the first assessment they complete and so do better in the second. While the partitioned groups are too small to be meaningful, comparing Stage and Luck Vogel's data by group (Fig. 11.3) shows Group A's correlations ($\rho = 0.59, p = 0.02$, $r = 0.71, p = 0.003$) are stronger and more significant than Group B's ($\rho = 0.45, p = 0.083$, $r = 0.6, p = 0.014$). This could indicate that the PEBL-first participants in Group A performed better in Stage because they are primed for these cognitive tests. It could also be that Group A just has stronger participants for this task (as we note P04 is a strong outlier who should not be removed as per App. B.4).



(a) Group A. (b) Group B.

Figure 11.3: Comparing Luck Vogel vs. Stage Rates by group.

Task Difficulty: The score distributions (Table 11.6) show us whether our measurement tasks are well-tuned. We expect a well-tuned task's score counts are normally distributed, potentially skewed to the right for our sample population. Comparing MGBatt to PEBL score distributions, we can infer their difficulty relative to each other in ways that could explain their (lack of) correlations. **Difficulty disparity between paired tasks could create correlation coefficients that are not good representations of the measures.**

Cake and Object Judgment: We see almost every participant received a perfect score — meaning the correlation calculation is difficult because the scores are almost constant. This suggests that Cake is too easy of a measure. In comparison, Object Judgment Absolute's distribution is so far from normal and fairly uniformly divided that we can tell it was hard to the point that participants may have been guessing. **This basic view tells us be wary of any correlations we see with Cake; and contextualizes why the Spearman for Cake-Object Invariant may be good, but the Pearson is so off.**

Looking, Recipe, Flanker and Four Choice: Looking and Recipe's normal-like distributions are a closer match to Four Choice's also normal-like distribution instead of Flanker's heavily skewed scores.

Stage and Luck-Vogel: Stage's distribution is steeper and more compressed than Luck-Vogel's distribution. As well Stage has lower scores. These elements suggest that Stage is much harder than Luck Vogel. This may be because of how Luck-Vogel presents the stimuli: in a central contained area so participants can more easily see all of them at once. In comparison, Stage's stimuli are larger, more animated, and spread out more across the screen. These attributes could be making Stage significantly harder than PEBL even though both are measuring the same construct.

Ability Isolation: Abilities are not mutually exclusive in task design or cognition, so measures incorporate some amount of error due to this overlap. **Weak correlations between tasks that should target the same ability may be the result of different supporting abilities**. For example, Looking is based on a choice reaction task, so we expect higher correlations with Four Choice Reaction Task. However, Looking asks participants to identify the target item in spite of distractors (similar looking items), where Four Choice only asks participants to select the quadrant where the stimulus appears. As well, Looking adds a non-trivial fifth choice — identifying if the target item is not a part of the presented set. This extra cognitive load from decision making could lead to weaker correlation. **Knowing we are unlikely to get strong correlations due to poor ability isolation, we are more likely to consider correlations greater than the critical values with significant p-values as targeting the same ability**.

11.4.2 Potential Latent Variable: Perceiving MGBatt as a Game

Participant Engagement: Participants seem more engaged with MGBatt; they were more likely to be leaning forward, talking to themselves, or nodding their head as they played the mini-games. Many commented on the music and sound effects being "catchy" and described mini-games as "fun". In comparison, participants describe PEBL as "frustrating" and "clinical". They look more serious and focused, sitting further back and not moving or talking to themselves in the same way. A couple of participants expressed that the mini-games felt like they were over quicker than the PEBL tests, and were surprised when we showed them the timings that indicated they were exactly the same. Participants would offer advice on how to make MGBatt more "fun" or "fair". For example, in the Looking and Recipe games which follow a Go/No-go structure, participants suggested that we constantly have the "go" condition displayed so they don't have to remember it. This would remove a core quality of the assessment, but would align more with commercial games and decrease difficulty. There were no suggestions offered as to how to improve PEBL, though participants were very vocal about the tasks they felt were frustrating (Object Judgment - Absolute condition).While we have no formal data on their perception of PEBL and MGBatt, the frequency of this kind of adhoc feedback could indicate that participants approached them differently. **The expectation that PEBL was "serious" and therefore something they needed to pay more attention could have skewed their performance higher for PEBL**

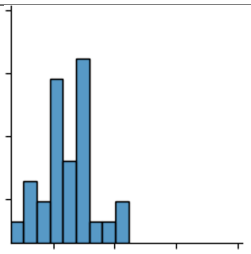
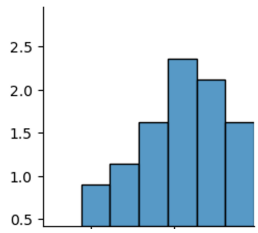
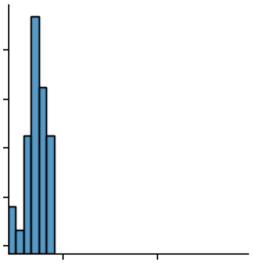
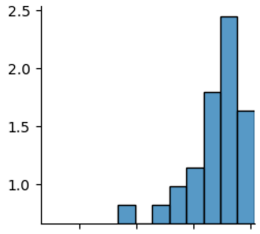
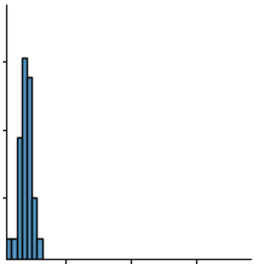
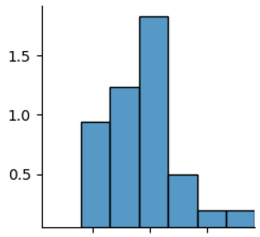
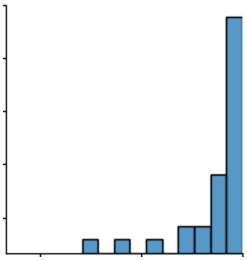
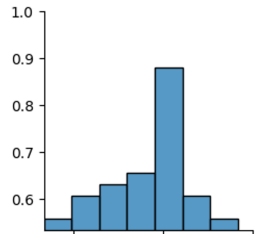
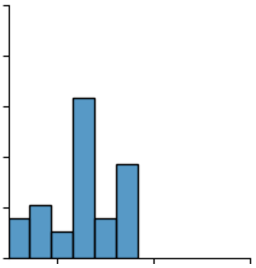
Battery Task	Score Distribution	PEBL Task	Score Distribution
Looking (Reasonable)		Four Choice (Reasonable)	
Recipe (Hard)		Flanker (Easy)	
Stage (Hard)		Luck Vogel (Reasonable)	
Cake (Easy)		Object Invariant (Good)	
		Object Absolute (Hard)	

Table 11.6: Score distributions for each Battery game and PEBL task. Distributions are labeled for their inferred difficulty tuning. Distributions are "good" if normally distributed, "reasonable" if they fit our expectation of normally distributed with slightly peaked and right-tail skew, "Easy" if they are exponential, and "Hard" if they are highly compressed.

tasks. Similarly the view of MGBatt as "entertaining" could have caused them to be more relaxed and error prone leading to worse performance.

Game Literacy: Perceiving MGBatt as a game could have led participants to believe they could rely on their game literacy to carry their performance. We notice that participants frequently skim or skip instructions to get to the game, only to realise they do not know the controls or what they are supposed to be doing. This happened in both the validity and reliability sessions, often resulting in input errors as participants would then press the wrong button (e.g. pressing Right shift as "no change" in Stage instead of Left shift) or forget options (e.g. forgetting that you can press 'X' to indicate the target item is not in the set for Looking). Participants were vocal about these mistakes both to themselves during play, and to us after their sessions. This kind of instruction skipping did not seem to happen with PEBL tests. Modern games often forgo tutorials and written instructions to allow participants to "figure out" the controls and how to play. **It is possible MGBatt's game-feel gave participants the impression they would have more space to practice or learn on-the-fly, and that the mini-games would provide more feedback and assistance to support their learning of the mechanics and controls.**

11.4.3 Controlled Variables

Fatigue: Back-to-back 20-minute assessments in the validation study means that the secondary measure could be affected by participant fatigue. We try to accommodate for this by counterbalancing the order via groups, and providing the the 5-minute break between segments. However we do not have a way to quantify the effects of fatigue or systematically account for it. This could make already difficult tasks significantly harder. As well, there could be fatigue inside each assessment as tasks are somewhat repetitive. MGBatt attempts to alleviate this by randomizing the game instances so you are not playing the same game consecutively; PEBL on the other hand presents all trials for a measure at once before moving to the next randomized task.

Task Switching: Since there are no breaks between the tasks inside a testing condition (PEBL, MGBatt) there may have been non-trivial task switching overhead. We attempt to account for this through task randomization. This would distribute task switching effects across the sample, which should reduce the effect of task switching on the overall data.

11.5 Conclusions from the Study

While we could only establish validity for Digger, our study results show promise for the validity of Looking, Recipe, Cake, and Stage. Despite the small sample size reducing study power and possible correlations, the fact that all the mini-games in MGBatt cross our significant correlation threshold ($\rho > 0.306, p > 0.05$) suggests that a larger study may validate the other games. **Given this interpretation, we think it seems feasible to use MGBatt games to measure abilities for a player profile.** However, we need to be cognizant of

the limitations of specific games (particularly Cake, Recipe, and Stage) due to their difficulty tuning, and ability overlap.

Future Work. Generally this work could be improved with a larger sample size, and a more diverse participant pool. Work looking to further develop this battery should focus on improving the MGBatt games, validating the games against other measures, and expanding the set of available games. We particularly discuss next steps for improving the games.

Better difficulty tuning: Games need to be better-tuned for measuring abilities. Difficulty needs to come from the ability-task relationship; consider Cake, the perceived difficulty comes from the speed of the conveyor belt (i.e. time-based stress) rather than the underlying ability. Improving its difficulty to really measure object recognition would require more complex categorization. However, the trade-off between "game-feel" and assessment difficulty needs to be further examined. For example, Looking's differences from a classic Four Choice task preserved its game-feel, but also introduced latent abilities that seem to obscure the relationship between the tasks.

Player Perceptions: As MGBatt is a game-based assessment tool it is important to capture that "game-feel". However, we did not anticipate the ways that it may lull participants into being more error prone. Future work should explore these implications further and see whether other groups replicate this trend. As well, going forward, these comparison measures should include more qualitative data about the ways participants engaged with each measure to see if there are any other latent variables to consider.

Take home points

From this chapter we learned the following meta-lessons:

- Our Mini-game Ability Battery games can be feasibly used to measure their associated player abilities.
- Rate scores are more reasonable and insightful than raw accuracy scores or reaction times for our measurements.
- The small, homogeneous group (mostly young, able-bodied, self-reported gamers) could be the reason we are seeing skewed data; future work would need to recreate this with a larger, more heterogeneous sample.
- The player's perception of the MGBatt as a "game" versus PEBL as a "test" may influence how they approach it (and then how they perform).
- We should have collected qualitative data about player perceptions to see if we could explain the results more.

Chapter 12

Closing Remarks: The Player Model

We end Part II with all the tools we need to make our Player Model. We develop the *cognitive and motor ability set* (Ch. 7 and 8) to describe players, and the *Mini-Game Ability Battery* (Ch. 10; some validation evidence in Ch. 11) to measure their abilities. We combine these into *player profiles*. Player profiles (Def. 3.2) visualize a specific Player Model instance. They are the quantitative model that will allow us to understand mechanical achievability.

We start this chapter by explaining how we construct player profiles from player data. We then explore the idea of a *player homunculus* — an imaginary player made from statistical data. We use the player homunculus and a specific player profile to show how we can analyse players against each other. We end this chapter with ways to improve player profiling, and a summary of this whole part.

12.1 Constructing Player Profile

Using P30 from the validation study as an example, we construct a profile in four steps:

1. Measure player abilities via Mini-Game Ability Battery.
2. Find the minimum value, mean, and standard deviation for sample scores.
3. Scale player scores.
4. Plot player profile.

We complete steps 1 and 2 in the validation study (Table 12.1).

		Sample Population	
Ability	**P30 Scores**	**Mean**	σ
Finger Pressing	90.09	77.62	12.09
Selective Attention	1.36	1.13	0.21
Inhibition	0.80	0.72	0.12
Object Recognition	1.00	0.95	0.08
Token Change Detection	0.24	0.20	0.05

Table 12.1: P30 Scores and Population Descriptive Statistics from Validation Study.

As per Design Decision 5, we normalize P30's ability scores to the same scale using min-max scaling, with the max and minimum values capped at $\pm 3\sigma$ from the mean. This allows

us to easily visualize and compare abilities, without losing the meaningful information from the ability units. For each ability, the scaled score is calculated as:

$$\text{Score}_{\text{Scaled}} = \frac{(\text{Score} - (\text{Mean} - 3\sigma))}{((\text{Mean} + 3\sigma) - (\text{Mean} - 3\sigma))}$$

We plot the scaled scores on a radar graph, and display the actual scores in callout boxes. We indicate motor abilities with a blue circle marker, and cognitive abilities with a purple diamond marker. Fig. 12.1 represents P30's specific player profile. Based on the markers' positions relative to the radar rings, we see P30 particularly is overall reasonably adept at these skills (close to the 0.75 ring).

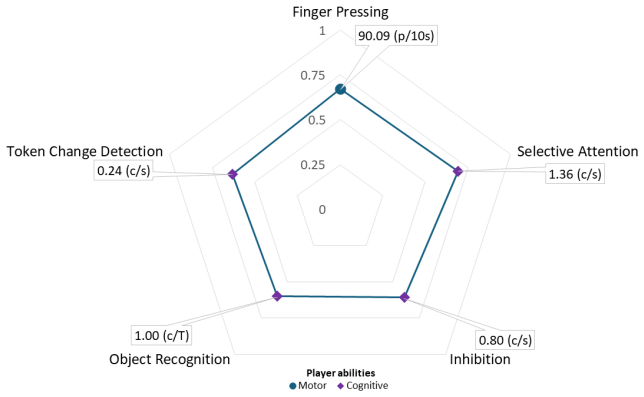


Figure 12.1: P30's Player Profile from Validation Study. ***Legend:*** *p: Button presses, c: Correct Responses, T: total number of trials, s: Seconds*

12.2 Player Homunculus

Player profiles can also represent imaginary players based on statistical distributions and values. We call these imaginary player profiles a *player homunculus*[1] (PH). We construct a PH in the same way we would a regular profile. For example, consider an imaginary "average player" from our validation study, whose abilities are set to the mean. We present their PH in Fig. 12.2.

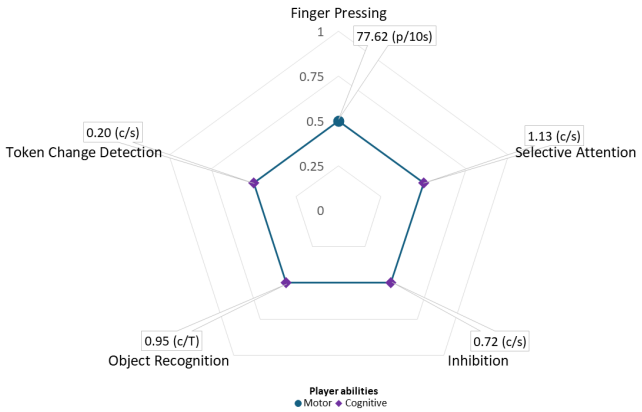
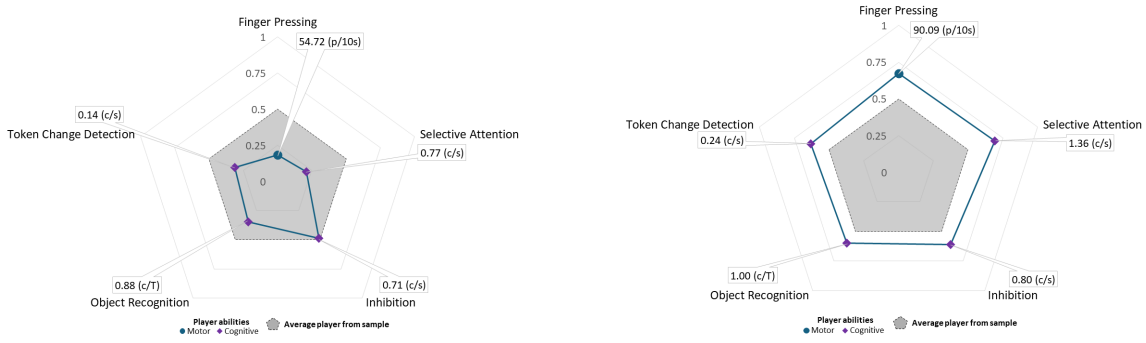


Figure 12.2: Player Homunculus: Average Player from Validation Study Population

[1] After the Motor Homunculus in Penfield and Rasmussen [370].

A PH can be useful for understanding players in relation to some persona, like the "average player". Imagine if we are designing a game and we tune its difficulty around the validation study's "Average Player" PH (Fig. 12.2). We can compare this PH to the profiles of P01 (Fig. 12.3a) and P30 (Fig. 12.3b) to get a quick idea of how significantly they differ in abilities. From this look, we see that P01 seems to have lower abilities than the PH. This could indicate they may struggle with the gameplay. In comparison P30 has higher abilities than the PH, which could indicate they will be able to handle the gameplay. This quick short hand allows us to make some hypotheses about the mechanical achievability of a game based on the mechanical achievability of the PH (calculating mechanical achievability is covered in Ch. 20).



(a) P01 (b) P30

Figure 12.3: Comparing Player Profiles from the Validation Study against the Player Homunculus of the "Average Player".

12.3 Improving Player Model

Future work on our ability-based player model should take two directions: expanding our existing measures, and gathering more data.

Expanding: Our current model only captures five abilities (1 motor, 4 cognitive). Expanding the Mini-Game Ability Battery to look at the other abilities in our set (20 motor, 21 cognitive) would enable us to have more complete player profiles. This expansion would need to be robust, and so should validate both the existing and new games against other standard measures. As well, the five abilities we currently capture should be further examined. As we note in Ch. 11, games capture multiple abilities. Therefore finding ways to combine measures that overlap in abilities, and adding new tasks that target these existing five abilities would allow for more redundancy and robust understanding of the player's actual ability levels.

More data: We currently have data about a small, fairly homogeneous sample of players. While this is sufficient for our thesis, our player model could be more useful if we have normative data for different player demographics. Currently we show how PH are useful for comparing a player to the sample average. Imagine how much more useful it could be if we

could construct PH for specific target demographics like neurotypical, able-bodied 10-year olds, or 17 year olds with cerebral palsy. Doing so would require larger studies focusing on finding normative data.

12.4 Wrapping Up

What we learned in this part:

- Existing models are inappropriate for our ability-based work because they are either too abstract (e.g. player typologies) or they are not specific to gaming contexts (e.g. GOMS).
- Our player model, inspired by the information processing model, focuses on basic cognitive and motor processes in a gaming context.
- Abilities (especially cognitive ones) are difficult (or virtually impossible) to isolate for measurements, so task designs often capture multiple abilities.
- Our game-based assessment of abilities are sufficient for approximately measuring player abilities.
- Tuning measurement task difficulty to find a person's ability limits is incredibly hard to do.

What we produced in this part:

- A motor model of 21 abilities (Tbl. 7.2);
- A cognitive model of 25 abilities (Tbl. 8.1);
- A battery of ability testing mini-games (Tbl. 10.2) implemented by two capstone groups and an MEng student under Sasha's supervision;
- Evidence of convergent validity for Digger, and suggested evidence of validity for other mini-games (Tbl 11.4) and a pilot set of reliability trends (Tbl. 11.5);
- A method for constructing player profiles (Ch. 12.1) from player data; and,
- The idea of the player homunculus (Ch. 12.2) for modeling imaginary players from statistical values.

Part III
The Challenge Model

We now cover the challenge model. We focus on modeling gameplay via *challenge competency profiles* (Def. 3.3). We originally conceived this model in our Masters work [442], and formalized it in our journal paper, "*Mechanical Experience, Competency Profiles, and Jutsu*"[441]. For completeness, we quickly review existing challenge frameworks, our definition of challenges, and our method for constructing challenge competency profiles, as described in these works (Ch. 13). We build from this starting point by: expanding the existing competency profiles to incorporate cognitive abilities from our player model (as per Design Decision 1), and experimentally validating our competency profiles.

We scope our work to focus on Button Mashing Challenges. We review the existing Button Mashing Challenge competency profiles and expand their cognitive abilities (Ch. 14). We design a series of studies to validate and explore button mashing competency profiles (Ch. 15). We validate the expanded competency profiles using multiple regression modeling (Ch. 16). We explore how these competency profiles change in responses to over and underloading limiting abilities (Ch. 17), and how player's perceptions change as well (Ch. 18). We end this part by presenting the final challenge models for Button Mashing games, reflecting on the lessons we learned about gameplay challenges, and highlighting potential next steps for this work.

Chapter 13

Background on Challenges

A Quick Aside...

The majority of this content has been presented in our paper: "*Mechanical Experience, Competency Profiles and Jutsu*" [441]. Minor additions and changes have been made for clarity.

Challenges are the unit tasks of gameplay, such that a full game is just a series of challenge [7, 115, 136, 293]. We are interested in understanding challenges in isolation to see how they build upon each other. To do that we need a working definition of a challenge.

The term "challenge" has been used to describe both gameplay activities and difficulty, which can make understanding its definition confusing. We focus on the use of challenge as a gameplay activity. Existing works from this angle (e.g. 7, 136, 505) differ on the specific definition of challenges, but agree that a challenge is an ability contest with win/loss conditions that are fundamentally characterized by a set of goals (i.e. what you are trying to do) and mechanics (i.e. how you can do it). We synthesize the various definitions into our working definition:

Definition 13.1. A *gameplay challenge* is any in-game activity with a success condition which engages the player in a way that requires some level of proficiency in at least one dimension (physical or cognitive).

Now that we have a working understanding of challenges, we can approach the problem of modeling them.

We begin by succinctly covering existing frameworks for describing challenges, from which we synthesize the components of a "good" challenge description (Def. 13.2). We then rehash our method for distilling atomic challenges, and analysing their required abilities. We illustrate this method with an example. We end this chapter by scoping the rest of our work to studying the Button Mashing family of challenges as a way to validate our method of challenge construction and the specific button mashing competency profiles.

13.1 Existing Challenge Frameworks

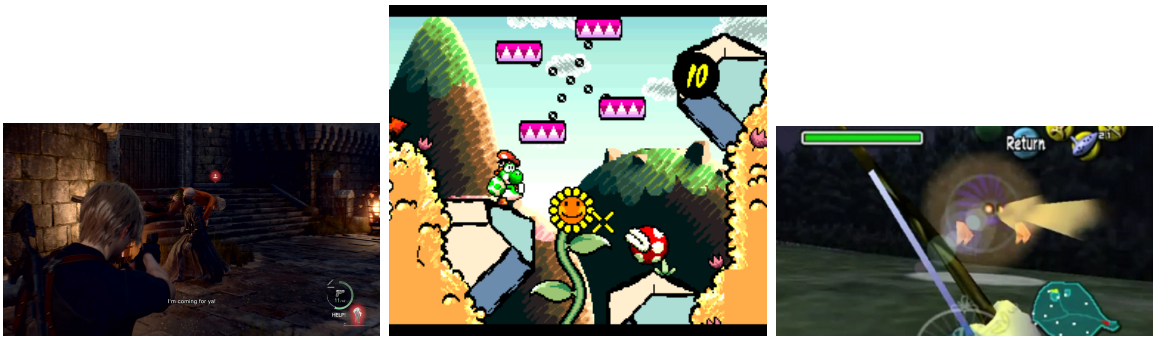
We look to identify *atomic challenges* — challenges that are mutually exclusive in their goals, context, and mechanical experience (MX). We found six frameworks for analysing gameplay and categorizing challenges [7, 47, 115, 136, 293, 505]. To decide if we can use one of these lists, we need criteria to judge their challenge descriptions. Ideally, a challenge description should include the in-game mechanics and the mechanism of interaction between the player and game (i.e. the inputs and outputs). The mechanics will let us understand the goals and actions of the challenge. The mechanisms of interaction provide the mechanical context and some insight into the MX. Adequate coverage of mechanics and mechanisms of interaction is only reached when we can differentiate between two similar challenges using the framework.

While we explore the works in the follow paragraphs, we can summarize the results easily. **All frameworks covered the in-game mechanics; none included the mechanisms of interaction. The coverage of in-game mechanics were insufficient to differentiate similar challenges for each framework.** We did not expect frameworks to discuss the mechanisms of interaction as generally work on understanding challenges tries to focus on the gameplay apart from the system context. However, we are surprised at the different ways mechanics and presentation were discussed.

Kinesthetic/Non-kinesthetic challenges. Veli-Matti views challenges as either kinesthetic (requiring non-trivial psycho-motor effort) or non-kinesthetic (requiring non-trivial cognitive effort) [505]. They further breakdown non-kinesthetic challenges as either static (i.e. puzzles, challenges with one singular solution) or dynamic (i.e. strategic, challenges with potentially many solutions). While this conceptual view reflects the nature of challenges as ability contests, it does not capture the mechanics to a sufficient level to differentiate challenges, nor does it capture the mechanisms of interactions to highlight the different types of motor and cognitive demands of the challenge.

Gameplay patterns. Bjork and Holopainen describe common gameplay patterns, combinations of which are analogous to challenges [47]. For example the pattern *aim & shoot* ("the act of taking aim at something and then shooting it") describes a common form of gameplay in first-person shooter games. While patterns do well at describing *what* is happening and, to an extent, the mechanics of the challenge, we do not get information about the mechanisms of interaction. Even though we get a sense of the mechanics of the gameplay through patterns, we cannot easily differentiate between similar challenges. Consider the *aim & shoot* pattern again; combining it with *guard* ("to hinder other players or game elements from accessing a particular area in the game or a particular game element"), *characters* ("abstract representations of persons in a game"), and *enemies* ("avatars and units that hinder the players trying to complete the goals") can describe gameplay as different as protecting Ashley in Resident 4 [76] (Fig. 13.1a), protecting Baby Mario in Yoshi's Island [347] (Fig. 13.1b), or defending Romani Ranch from aliens in The Legend of Zelda: Majora's Mask [331] (Fig. 13.1c).

PLAY, GAME, META Bricks. Djaouti et al. create a system of "bricks" whose combinations describe gameplay [115]. The "PLAY" bricks describe player actions, "GAME" bricks describe goals, and "META" bricks are the combinations of PLAY and GAME bricks that



(a) Resident Evil 4. (b) Yoshi's Island. (c) Majora's Mask.

Figure 13.1: Examples of different gameplay that all use the aim & shoot pattern.

create families of challenges. While the mechanics are covered, mechanisms of interactions are not. As well, we run into issues of differentiating between similar challenges.

Feil and Scattergood [136] Challenge Taxonomy. Feil and Scattergood identify six challenges: time, dexterity, endurance, memory/knowledge, cleverness/logic, and resource control [136]. They define these challenges broadly, making it difficult to understand the underlying mechanics. This generally makes it difficult to differentiate between similar challenges. There is no discussion about the mechanisms of interactions.

Adams [7] Challenge Taxonomy. Adams defines ten major challenges, subdivided into thirty specific ones (Table 13.1). The specific challenges are not clearly defined, often relying instead on a series of examples to characterise the challenge's gameplay. The examples given show how broad many of the challenges are; meaning we cannot easily differentiate between similar challenges, since diverse ones are often lumped.

McMahon, Wyeth, and Johnson [293] Refined Taxonomy. McMahon, Wyeth, and Johnson refine Adams taxonomy through an expert focus group, leading to only sixteen challenges. The refinement grouped together multiple types of challenges, and in our opinion, has made it more difficult to see each challenge type as describing a singular type of gameplay. Nuance about presentation and mechanics are not covered in this refinement. There is no discussion of the mechanism of interactions.

13.1.1 What is a good challenge description?

Having not found a suitable existing framework, we create our own taxonomy of atomic challenges. We use Adams [7] as a starting point because it understands challenges as having physical and cognitive requirements, and is grounded in gameplay examples. This reflects our ability modeling goals, and makes it easier to figure out what Adams intended for the challenges and how we can expand on them.

Challenge Type	Challenges
Physical Coordination	Speed and Reaction Time Accuracy and Precision Timing and Rhythm Learning Combination Moves
Formal Logic	Deduction and Decoding
Pattern Recognition	Static Patterns Patterns of Movement and Change
Time Pressure	Beating the Clock Achieving something before someone else
Memory and Knowledge	Trivia Recollection of objects or patterns
Exploration Challenges	Identifying spatial relationships Finding keys (unlocking any space) Finding hidden passages Mazes and Illogical spaces
Conflict	Strategy, tactics, and logistics Survival Reduction of enemy forces Defending vulnerable items or units Stealth
Economic	Accumulating resources or points (growth) Establishing efficient production systems Achieving balance or stability in a system Caring for living things
Conceptual Reasoning	Sifting clues from red herrings Detecting hidden meanings Understanding social relationships Lateral thinking
Creation and Construction	Aesthetic success(beauty or elegance) Construction with a functional goal

Table 13.1: Gameplay Challenges from Adams [7, p. 19-20]

Before we dive into making our own list of atomic challenges, we want to be clear on what our model (i.e. challenge description) must cover.

Definition 13.2. A good challenge description must delineate between similar challenges, and so includes the following:

1. the in-game mechanics associated to the challenge,
2. the mechanism of interaction between the player and the game, and
3. the intrinsic competency profile (i.e. the particular cognitive and motor abilities used to complete the challenge).

13.2 Refining Adams to Atomic Challenges

Recall Design Decision 4: challenges are distinct when they have different abilities or observable differences in their ability requirements. With this in mind, we refine Adams taxonomy through an iterative process.

1. Clarify Adams: Adams challenges are presented inconsistently; some explain the mechanics, others are explained solely through examples. We look through his work for clues about the experiences he is describing. The major challenge type hints at whether the challenge is more physical or cognitive. Any written definitions, if provided, give us insight into the mechanics and sometimes a mechanism of interaction. We play and observe others play any examples and document similarities between them to try and approximate the challenge. Through this process we arrive at a consistent, specific, definition for Adams' challenges. However, this definition is not yet precise enough.

2. Find Support: We search across genres and systems to find other gameplay instances that fit our new definition. This gives us insight into how a particular challenge can present in different contexts. There is no easy way to systematically search for this information, so we rely on our subjective knowledge of games to lead this search. Our collective gaming experience[1] spans more than 20 years, covering the third to eighth generations of home consoles, arcade games, and home computers from MS-DOS to Windows 10, and a variety of game genres. This part of the process would benefit from a larger pool of researchers with different gameplaying experiences, but is a sufficient starting point.

3. Interaction Grouping: We sort the examples by *mechanism of interaction*, as this is the most easily identifiable difference. We understand the *mechanism of interaction* by examining the game mechanics, instructions, and controller for the instance. This first separation accounts for obvious differences in motor abilities used, even if the abstract goals are the same. We also note down and separate examples by game mechanic variants, like pressing two buttons at a time instead of one button.

4. Close Reading: We play the examples multiple times, and use close reading techniques[2] (e.g. 46) to report on the details of our play experience. We take the role of naïve player, and attempt to play through each instance as if it is our first time. In our first play through we attempt to become familiar with the game mechanics, and note down our original impressions of the gameplay. We then repeat the gameplay, and reflect on the particular abilities we believe are being used in play. We continue this kind of naïve play mentality and experiment with our interactions with the game to try and understand how the mechanics respond to our different abilities. We systematically observe our performance and rank our use of each ability as it relates to our completion of the challenge (ranks in Table 13.2). We then separate

[1] Myself (Sasha), Dr. Jacques Carette (supervisor), and peers in the G-ScalE lab

[2] We note the close reading of games can be affected by the game's difficulty [46], but as we are particularly trying to read information about the game difficulty we believe this is not a major problem.

the examples into groups that all use the same abilities in the same order; these become the new challenges we are examining.

Rank	Range
Not used	0
Used, but not noticeable	1-25
Noticeably used	26 - 50
Important but not limiting	51 - 80
Limiting	81 - 100

Table 13.2: Ability rankings and their value ranges.

5. Repeat: We repeat steps 1 to 5 until we begin to see the examples stabilize into a finite number of categories. Once stabilized, we compare the examples inside each stabilized category to each other. We focus on the context of each example, like whether the game is competitive or co-operative, single or multiplayer, team-based or solo, etc. We try to see if these contexts create any differences in the way we have assessed the abilities. If there are notable differences, we subdivide the category again and repeat the process. **At the end of this iterative process we should have groups of challenges that are notably distinct in their abilities.**

6. Hypothesize: For each group, we re-examine the examples and assign each ability a value between 0 and 100 to represent how much we think it is used in the challenge (with a margin of error of ± 10). While unrealistically precise, this helps us express the finer differences between abilities — especially those in the same rank or on the borders. While this remains subjective, we have sample-tested our assignments against others' subjective classification (within our lab), and found our rough numbers to be uncontroversial. **These refined ability values become the competency profile for the group.**

13.2.1 Refinement Example: Speed Challenges

We split Adams' *Speed and Reaction Time* challenges, and apply our process to Speed Challenges in detail. This illustrates how our refinement approach leads to new distinct categories, each with simpler descriptions. Currently our descriptions only concern an individual's mechanical experience of these challenges. Thus, while we include examples of multi-player games, we examine them when playing with or against humans, or non-player characters.

Clarify: Speed challenges "test the player's ability to make rapid inputs on the control" [7, p.262]. As *Physical Coordination* challenges they are motor-focused. Describing them as "rapid" indicates a *time limit*; "inputs on the controls" implies controller-independent gameplay. So examples should exist using all controller types. There is no mention of particular stimulus that would trigger this action. This is likely due to the distinction between Speed challenges and Reaction Time challenges, where the latter relies on a specific stimulus for a "reaction". So gameplay instances that require players to "react" and not just "act"

do not belong to Speed Challenges. Furthermore, these challenges should be identifiable as small chunks of gameplay – not something that takes place over the course of hours of a play session. Adams lists Tetris [9], Track & Field [238], and Quake [203] as examples, without giving specific instances inside these games to pinpoint what he means. He does list platformers, shooters, and fast puzzle games as genres where these are most readily found. Deeper analysis reveals these are more instances of reaction time over speed challenges. **We summarize this challenge type as exclusively motor-focused, and defined by short sessions and time limits**.

Support: We find several examples of short session, time limited, and motor-focused gameplay in party games and mini-games. Nintendo games are particularly popular in this genre and exist across multiple input mechanisms; this gave us ten examples:

1. Manic Mallets, Mario Party 5 [197];
2. Cycling, Mario and Sonic at the Olympic Games [420];
3. Mecha-Marthon, Mario Party 2 [194];
4. Pedal Power, Mario Party [193];
5. Tenderize the Meat, Cooking Mama [95];
6. Impressionism, WarioWare: Touched! [208];
7. Wash Rice, Cooking Mama [95];
8. Hammer Throw, Mario and Sonic at the Rio 2016 Olympic Games [421];
9. Candy Shakedown, Super Mario Party [324];
10. Trike Harder, Super Mario Party [324].

Interaction Grouping: These gameplay instances have different mechanisms of interaction, implying different underlying motor abilities. We group these challenge into: *button mashing, rapid analog stick rotation, rapid tapping, scribbling, rapid controller rotation, and rapid controller shaking*. We now begin the iterative process of refining these, starting with button mashing.

13.2.2 Refining Further: Button Mashing

Clarify: Button Mashing is where a player must rapidly press button(s) or key(s) in a given time limit. From our original list button mashing appears in:

1. Manic Mallets, Mario Party 5 [197];
2. Mecha-Marthon, Mario Party 2 [194];
3. Track & Field [238].

Support: We easily find more (Nintendo) instances. The abundance of examples argues that this is a common category of challenge in the party and mini-game genres. Expanding outside Nintendo and party games is more difficult as instances tend to be embedded in larger gameplay segments. Overall we list eleven additional examples (seven Nintendo, four non-Nintendo):

1. Psychic Safari, Mario Party 2 [194];

2. Speed Skating, Mario and Sonic at the Winter Olympic Games [423];
3. Ridiculous Relay, Mario Party 3 [195];
4. Take a Breather, Mario Party 4 [196];
5. Pump, Pump, and Away, Mario Party 3 [195];
6. Chin Up Champ, Wii Party [348];
7. Balloon Burst, Mario Party [193] and Mario Party 2 [194];
8. Torture Attacks, Bayonetta [375] and Bayonetta 2 [376];
9. Dragon's Breath, South Park: The Stick of Truth [355];
10. Boss Knockouts, Donkey Kong Coountry: Tropical Freeze [391];
11. Colossus of Rhodes Fight, God of War 2 [412].

Close Reading: Across instances button mashing is hardware dependent, requiring physical controls to depress (ergo "buttons" to "mash"). This is different than pressing virtual buttons like those found on a touch screen as it loses the mechanical feedback of a button. Therefore we will not need to further group these by interaction. The time limit can be either implicit (often being tied to the length of an animation or just not explicitly shown to the player) or explicit (timers or gauges). However, we did not find that explicit versus implicit time limits affected our mechanical experience. Generally, we were too focused on pressing quickly to watch the timer when it was explicit. As well, since the goal in every instance is to press the buttons as quickly as possible, there was no change in our play style or strategy. This is likely because of the simplicity of this particular challenge; we believe explicit time limits would affect cognitive-focused challenges more.

Grouping: While similar, do these examples have the same game mechanics? Consider Manic Mallets [197], Speed Skating [423], and Mecha-Marathon [194]. In Manic Mallets the player hits a single button as many times as possible in the time limit; Speed Skating requires the player alternate between two buttons; Mecha-Marathon requires pressing two buttons simultaneously as many times as possible. Manic mallets, with its single button, is a straightforward case of button mashing, requiring no additional abilities outside of pressing the button. Mecha-Marathon requires some coordination of button pressing, principally focusing on the pressing but requiring some attention. Speed Skating similarly requires finger pressing and attention, adding a small perception and memory component to keep the alternating pattern correct. All the other examples repeat one of these three patterns. **These differences in used abilities divides button mashing based on type of input: *single, multiple, and alternating.*** We refine these further in Ch. 14.

13.3 Scoping Our Work

Refining Adams' complete list of challenges is still a work-in-progress. While the Mechanical Experience paper [441] continues to further refine and explain the different types of button mashing, we take a moment here to pull back and talk about the scope of our thesis challenge model.

The goal of this part of the thesis, is to show that these profiles are valid, and by extension this method of generating them is reasonable. As such, we want to scope the rest of this

part of the thesis to challenges we can experimentally validate. Generally we think our work could explore one of three broad types of challenges: motor-limited, cognitive-limited, and mixed challenges. Given the existing length of this thesis, and the time constraints for its completion, it is unfeasible to generate and validate challenge competency profile for each broad type. Therefore we need to scope our work to focus on one type.

13.3.1 Challenge Selection Criteria

We want the simplest version of a challenge type, meaning the fewest number of abilities to test and the most apparent difference in ability usage. An ideal challenge then has one very obvious limiting ability (highly correlated to performance), with all other ability types being in the used, but unimportant or lower range (low to no effect on performance). **We believe the simplest case to test is a motor-focused challenge.**

Motor-focused Scoping. The ideal motor focused challenge should have a singular limiting motor ability, and little to no cognitive load. This would create a large difference between the limiting motor ability and other abilities, making it easier to test the competency profile. The motor ability should be simple, as complex movements make finding the limiting motor ability hard. For example, in a game of Dance Central for the Kinect it's difficult to tell whether the arm movements or leg movements contribute more to your dance score. This gets even more complicated when limited to a particular body part – for example, in Mario and Sonic at the Olympic Games Tokyo 2020 [419], when doing the javelin throw players must shake the Joycon to build up power as they run and then adjust the angle of their throw with their wrist movements before launch. The shaking motion alone may consist of shoulder and forearm movements depending on the player; differentiating the limiting ability between shoulder, forearm, and wrist is then more complicated as we'd need to figure out how much each segment of the javelin mini-game contributes to the final score.

Constraint. As we note in Ch. 10.1.1, COVID-19 constrains our experiments to challenges that could easily be performed on home computers or laptops[3]. So our motor ability needs to relate to keyboard inputs (e.g. finger pressing).

Considering our ideal challenge criteria and constraints, we think the best motor-focused challenge would be the family of Button Mashing challenges. Button mashing challenges have a single motor ability focus, and because they are speed challenges their cognitive load seems to only rely on sensory perception (recognizing stimuli) and limited attention (for focusing on the game).

13.4 Summarizing Background on Challenges

With a clear scope and refinement method, we can move forward with developing our challenge model. Over the next chapters we must:

[3]The design and creation of the experiment and its apparatus was done during the COVID-19 pandemic. However, in-person testing was possible by the time we had completed the designs and received ethics approval. As such we leave this rationale to reflect the thought process at the time.

- Clearly summarize the button mashing challenges and their competency profiles; and,
- Design and execute a study to validate their competency profile.

Take home points

From this chapter we learned the following meta-lessons:

- We review 6 existing challenge frameworks, and decide to focus on refining Adams taxonomy of challenges [7].
- We define the elements of a "good" challenge description based on its ability to differentiate gameplay.
- We outline an iterative method for refining Adams challenges into more precise challenge descriptions.
- We scope our work to Button Mashing challenges.

Chapter 14

Button Mashing

Button mashing is a speed challenge that comes in three types: *single input*, *alternating input*, and *multiple input*[1]. Soraine and Carette [441] refine and analyse these challenges focusing on their motor abilities. We now focus on expanding their cognitive elements.

For each challenge type, we summarize the original refinement and motor analysis results. We then present the new expanded competency profiles. We support these results by summarizing our close readings of gameplay examples. Our analyses focus on the individual player's experience. So while some of our examples are multiplayer games (ergo "success" may be inter-player dependent), we focus on the abilities and mechanics of singular players.

A Quick Aside...

To help understand our readings, we provide links to ability definitions and YouTube videos of gameplay as framed images.

14.1 Single Input Button Mashing (SIBM)

SIBM are motor focused challenges where players "repeatedly press a specific single button or key as fast as possible within a given time limit" [441]. SIBM are frequently a sub-challenge of larger gameplay, but can be found as the main gameplay in Party Games like the Mario Party series. Their main mechanics are: button pressing, and short time limits. Tunable variables for these mechanics are: the button being pressed, the length of the time limit, and what "counts" as a press. Fig. 14.1 summarizes our initial competency profile for a single player competitive context on a standard controller. Finger pressing is the limiting ability (as it determines how many presses you can make in the time limit) and some perceptional (for knowing what button to press) and attentional (for focusing on the game) abilities are used in a minimal capacity. The original gameplay instances used to create this competency profile are:

- Manic Mallets, Mario Party 5 [197];
- Dragon's Breath attack, South Park: The Stick of Truth [355];

[1]These types were refined from Adams [7] Speed and Reaction Time Challenges

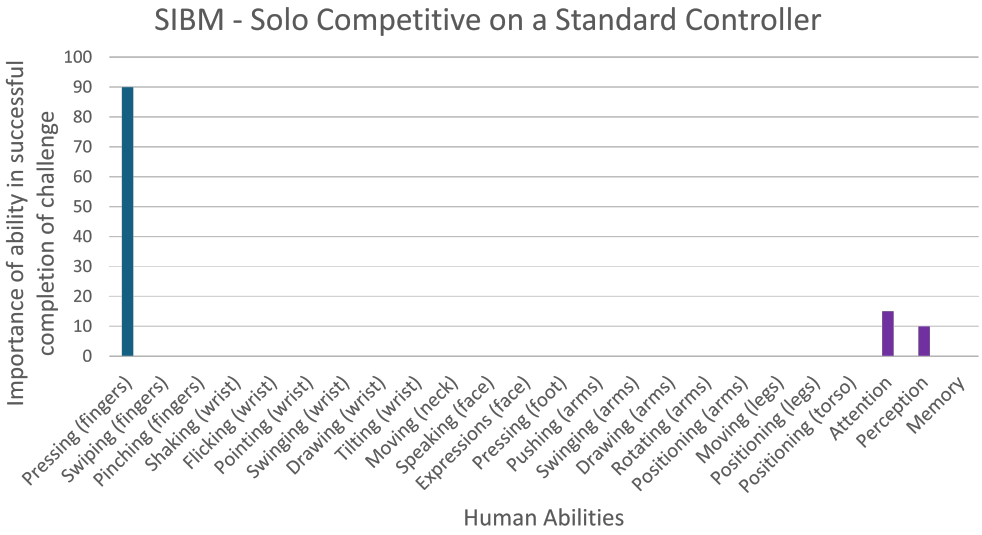


Figure 14.1: Competency Profile for Single Input Button Mashing as played on a standard controller from Soraine and Carette [441]

- Torture attacks, Bayonetta [375] and Bayonetta 2 [376]; and,
- Boss Knockouts, Donkey Kong: Tropical Freeze [391].

We re-examine these SIBM examples using our game reading techniques to add cognitive detail. We identify the following abilities, ordered by importance, as part of SIBM:

1. Finger pressing (Score: 90, Rank: Limiting),
2. Selective attention (Score: 15, Rank: Used),
3. Inhibition (Score: 15, Rank: Used),
4. Object recognition (Score: 10, Rank: Used),
5. Tactile perception (Score: 10, Rank: Used), and
6. Procedural memory (Score: 10, Rank: Used).

Recall from the design of our player profiling tool (Mini-game Ability Battery), that tactile perception and procedural memory are captured as part of the measurements for the other abilities. Since we aim to validate these profiles (Ch. 16) we choose to reduce the list to measurable abilities. Fig. 14.2 graphically represents this reduced competency profile, showing each abilities estimated use (with a ± 10 error) and rank.

14.1.1 Summary of Gameplay and Ability Close Readings

We consider Manic Mallets to be a prototypical SIBM, and so favour its distribution when considering the abstract competency profile for this challenge type. Fig. 14.2 is a synthesis of the information below, so we do not repeat it.

Manic Mallets [197] has teams of two players repeatedly hitting a switch with a hammer to avoid being crushed by a bigger hammer (Fig. 14.3). Pressing A swings our little hammer once. Each player must press A as fast as possible (therefore as many times as possible) in the ten second time limit of the game; the team with the most cumulative presses wins.

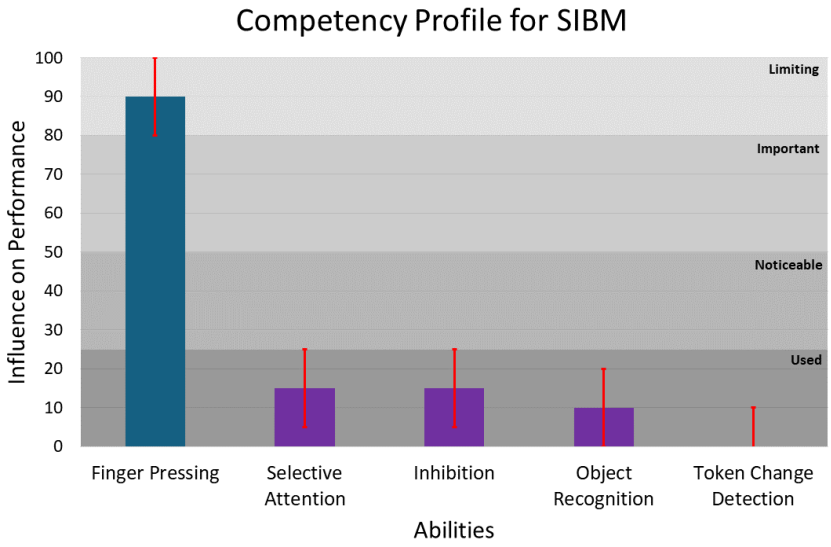


Figure 14.2: SIBM Competency Profile Hypothesis. Blue: Motor, Purple: Cognitive.

Figure 14.3: Manic Mallets (Mario Party 5) by NintendoMovies.

Finger pressing is the most important ability (Score: 90, Rank: 4), as the speed and repetition needed for success means that any issues with that ability would greatly affect performance. **Selective attention** and **inhibition** are somewhat important at keeping us focused on our performance, but not impediments to success (Score: 15, Rank:1). The big hammer provides visual performance feedback; if it is pointing towards our character we are doing poorly, but if it is pointing at our opponents we are doing well. Paying attention to the hammer helps us decide whether we need to change the speed, intensity, or even our hold on the controller (all of which can influence our button throughput). However, these attention abilities are less important because Manic Mallets is a stand-alone mini-game and so we do not need to inhibit pressing after the time limit (so we do not need to pay *as much* attention). We also use **object recognition** to know which button we need to press, and **tactile perception**[2] to know when it is pressed. This is important because there are no visual or audio stimuli reflecting every individual button press. Tactile perception is supported by **procedural memory** to reorient ourselves on the controller if our fingers slip. Object recognition, tactile perception, and procedural memory are used (Score: 10, Rank: 1) in the playing of the game, but are largely unimportant to the player's overall success.

The Dragon's Breath Attack in South Park: The Stick of Truth [355] is a mage class attack where player's wave a lit firecracker in front of the enemy NPCs face do deal combat

[2]In the case of the Nintendo Gamecube controller (where the A button is larger than the other buttons), tactile perception also supports object recognition since we can tell which button it is by feel.

damage (Fig. 14.4). This game is available on multiple platforms affording different inputs/-control schemes; for this instance we are looking at it on a standard controller. The control scheme is presented by the game as "mash[3] the A button". The time limit of this challenge is tied to the length of the attack animation which lasts for about 3 seconds.

Figure 14.4: Dragon's Breath attack (South Park: The Stick of Truth) by The Game Cavern.

Finger pressing is the motor interaction, and limiting factor in play (Score: 90, Rank: 4). Poor motor control will slow down pressing which directly results in poor performance. Similar to Manic Mallets, **object recognition** and **tactile perception** underlie knowing what to press and if we have pressed it. **Procedural memory** becomes more important as the button layout changes depending on the standard controller used (e.g. Xbox One, Nintendo Switch Pro Controllers, and PlayStation DualShocks). Overall these are used but not important (Score: 10, Rank: 1). **Selective attention** and **inhibition** keep us focused on the gameplay for the time limit. Since this SIBM instance is a small part of a larger combat system, the player must focus so as to stop at the appropriate time. Continuing to mash after the animation time registers as menu inputs for another part of combat, making stopping on time crucial to effectively engaging in combat. On reflection, when playing there was enough time around the start and ending of the challenge that it was trivial to focus on when pressing the button is useful, so we decided they are used, bordering on noticeable (Score: 20, Rank: 1).

Torture Attacks from the Bayonetta series [375, 376] are a triggered[4] combat action which removes players from regular combat to perform a "quick time event" (QTE) to increase their score and deliver a cinematic finishing blow. The QTE challenges are mostly button mashing segments[5] (see Fig. 14.5). This game is released on multiple platforms; for this analysis we considered the standard controller as the input device. When activated, the torture attack displays a circular gauge with a button on it corresponding to the button on the controller you must press. Therefore the control scheme is visible at all times. The time

Figure 14.5: All Torture Attacks (Bayonetta) by Catan.

[3]Mash is interpreted to mean "quickly and repeatedly press".
[4]Players initiate torture attacks by pressing Y and B when the signifier appears on screen.
[5]One outlier requires the player to rapidly spin their left thumbstick

limit is implicit and tied to the length of the attack animation.

Finger pressing is still the limiting ability (Score: 90, Rank: 4), as poor motor performance has the greatest impact on overall performance in this segment. **Object recognition** is used to know what to press; this is especially obvious as the game displays the buttons to press throughout the torture attack. **Tactile perception** and **procedural memory** are also used to recognize whether a press has happened, and to orient ourselves around the controller. We understand these abilities are used because of the nature of the task, but the context of this instance (displayed inputs, outside of regular combat) makes them used but not important (Score: 5, Rank: 1). **Selective attention** and **inhibition** are noticeably used abilities (Score: 20, Rank: 1). Like the Dragon's Breath attack, they are slightly more important here than in other SIBM challenge due to existing inside the larger combat system of the game, and being integrated into the normal gameplay.

Figure 14.6: Boss Knockout of Skowl (Donkey Kong Country: Tropical Freeze) by BossBattleChannel

Boss Knockouts in Donkey Kong Country: Tropical Freeze [391] are boss-fight specific QTEs[6] where the player mashes a button (X or Y) to deliver "finishing blows" to the boss, thus increasing their final score for the level. As this segment is outside of regular gameplay, and after the boss fight gameplay, it is impossible for the player to "fail" this challenge — rather success here is about achieving the high score. This game is released on both the Wii U and Switch, meaning that there are multiple control schemes that influence which button is pressed. However as it is always a button press, we focus on the Wii U version with a Pro Controller (standard controller) as it is representative of all instances. During the Boss Knockouts, the button to be pressed is shown on the top right corner of the screen. The challenge timer is not displayed on the screen, but every instance lasts for about four seconds.

Finger pressing enacts the punches and remains the limiting factor in performance (Score: 90, Rank: 4); poor motor ability leads to fewer punches and a lower score. **Selective attention** and **inhibition** are noticeably used (Score: 15, Rank: 1) but less important here than our previous two examples because the instance is a separate gameplay mode (with its own camera model and control scheme, separate from the normal gameplay and boss fights) that happens at the end of the level. As such stopping late is meaningless, and pressing the wrong button has minimal effect on the player's overall score. **Object recognition**, **tactile perception** and **procedural memory** are abilities used to ensure we are pressing the correct button during this gameplay segment. As the button we need to press is on screen for the duration of this segment, and there is no significant penalty for pressing the wrong button, we believe that these abilities are used but not important (Score: 10, Rank: 1).

[6]Triggered when boss's health is reduced to zero (i.e. the player won the fight)

14.2 Alternating Input Button Mashing (AIBM)

AIBM are motor focused challenges where players "repeatedly and rapidly pressing two or more specific buttons in sequence" [441]. Like SIBM, they are predominantly found as part of larger gameplay challenges, or independently in mini-games. Their main mechanics are: button pressing, short time limits, and pattern recognition. The main variables are the length of the sequence (which increases how much a player needs to remember), which buttons are being pressed (as their distance may contribute to physical pressing difficulty and/or input errors), what "counts" as a button press, and the time limit.

Fig. 14.7 summarizes our initial competency profile for a multiplayer competitive context on a standard controller. The limiting ability is finger pressing. Attentional abilities are significantly more important in this challenge than other button mashing instances in order to maintain input sequencing.

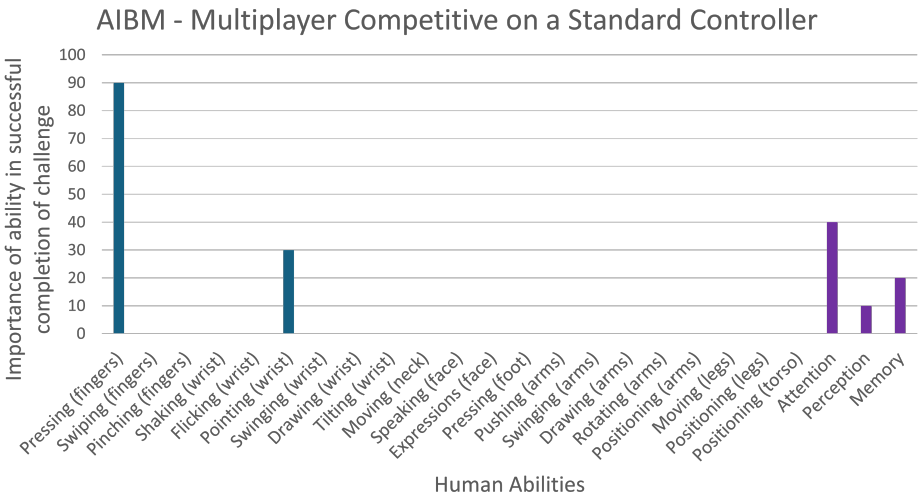


Figure 14.7: Challenge Description for Alternating Input Button Mashing as played on a standard controller from [441]

The original gameplay instances used to create this competency profile are:

- Psychic Safari, Mario Party 2 [194];
- Balloon Burst, Mario Party [193], Mario Party 2 [194], and Mario Party Superstars [323];
- Pump, Pump, and Away, Mario Party 3 [195];
- Ridiculous Relay, Mario Party 3 [195];
- Speed Skating, Mario and Sonic DS [420]; and,
- the Colossus of Rhodes fight, God of War 3 [413].

We revisit the gameplay instances and identify AIBM's complete competency profiles as:

1. Finger pressing (Score: 90, Rank: Limiting),
2. Inhibition (Score: 40, Rank: Noticeable),
3. Selective attention (Score: 40, Rank: Noticeable),

4. Token change detection (Score: 20, Rank: Used),
5. Object recognition (Score: 10, Rank: Used),
6. Tactile perception (Score: 10, Rank: Used),
7. Procedural memory (Score: 10, Rank: Used)

Following the same logic as SIBM, we reduce these to measurable abilities for future validation purposes, arriving at the reduced competency profile in Fig. 14.8.

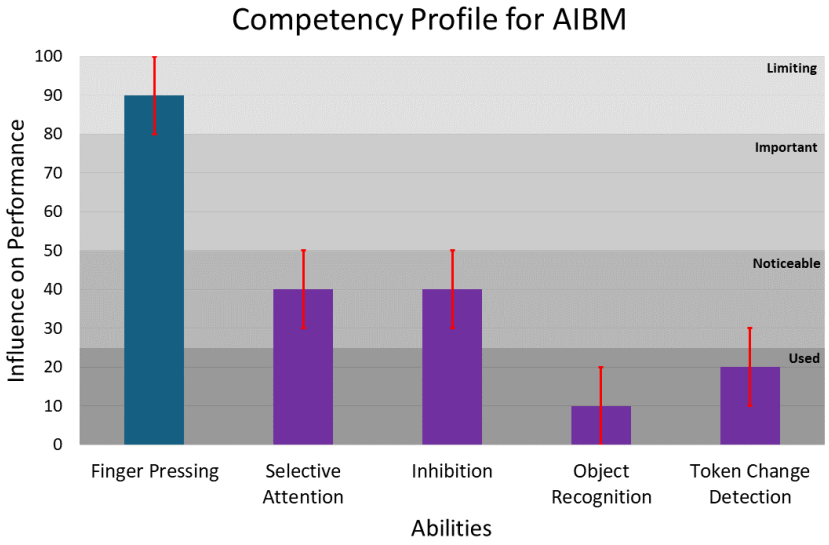


Figure 14.8: Hypothesized AIBM Competency Profile. Motor abilities are Blue, Cognitive abilities are Purple.

14.2.1 Gameplay-Ability Descriptions

We view Psychic Safari as the prototypical example of this challenge, and so favour its distribution in synthesizing the above competency profile.

A Quick Aside...

During our review, we found that **Balloon Burst** and **Pump, Pump and Away** do not fit the button mashing family. While button mashing is a reliable ***strategy*** versus Easy and Normal CPU opponents, it was no longer effective at the Hard and Super Hard levels where a well-timed sequence was more competitive[a]. This implies these games are based more on timing, with the pumps visually indicating the "correct" pace to make inputs. Our original analyses only looked at Normal level CPUs and against player opponents in various configurations (e.g. human-vs-npc, human-vs-human) [441] and so missed this game dynamic.

[a] Against human players the effectiveness of button mashing versus timing is a toss-up, depending on the techniques your opponents use.

Psychic Safari [194] asks players to power up an ancient relic by alternating between pressing the A and B buttons (Fig. 14.9). The game displays which button in the sequence the player currently needs to press above their character. There is an explicit 5 second time limit and the player who can make the most inputs wins (and destroys the other player's relic).

Figure 14.9: Psychic Safari (Mario Party 2) by NintendoMovies.

Finger pressing is the most important ability (Score: 90, Rank: 4). The player's button pressing paces are mutually exclusive, and it is a race, so their individual speed is the biggest determinant of success. **Selective attention** (visual) and **inhibition** are more important in this game because of the alternating pattern. Players lose a lot of time and momentum in this challenge if they end up pressing the wrong button. However, the short duration (5 seconds), short repeating sequence (A-B), and visual display of the pattern means that these attention abilities are supported by perception as well and so may be noticeable, but not excessively important (Score: 40, Rank: 2). **Object recognition** (Score: 10, Rank: 1) is used to recognize which button to press, and **token change detection** (Score: 20, Rank: 1) is used to recognize the visuals switching between buttons during the gameplay. Since the button that needs to be pressed is constantly on display, token change detection seems to be more important because it clues the player into the sequence and so supports getting "back on track" should their inhibition fail and a mistake was made. **Tactile perception** and **procedural memory** support play through identifying whether the button is sufficiently pressed and the orientation of the controls, making them used but not noticeable compared to the other abilities (Score: 10, Rank: 1). Procedural memory may also help with the patterned pressing; however, with such a short sequence accompanied by visual reminders, this ability is overall less important.

Ridiculous Relay [195] is a 1-vs-3 player mini-game, where the solo-player races against the three-player tag-team. The catch in this mini-game is that all the players have a different control schemes. We focus on the first member of the three player team as an example of AIBM[7]. The first teammate (Fig. 14.10) must paddle their boat by alternating pressing A and B (corresponding to paddling left and right). The game displays which button the player currently needs to press above the character. The game does not have a time limit;

[7]Fun fact: the third member (last stretch of the relay) is an example of SIBM

instead the middle of the screen is divided with an abstracted view of the race course to see the players' progress and gauge who is winning. "Success" for the AIBM section is fastest completion of the first segment, so effectively the time limit is self-determined.

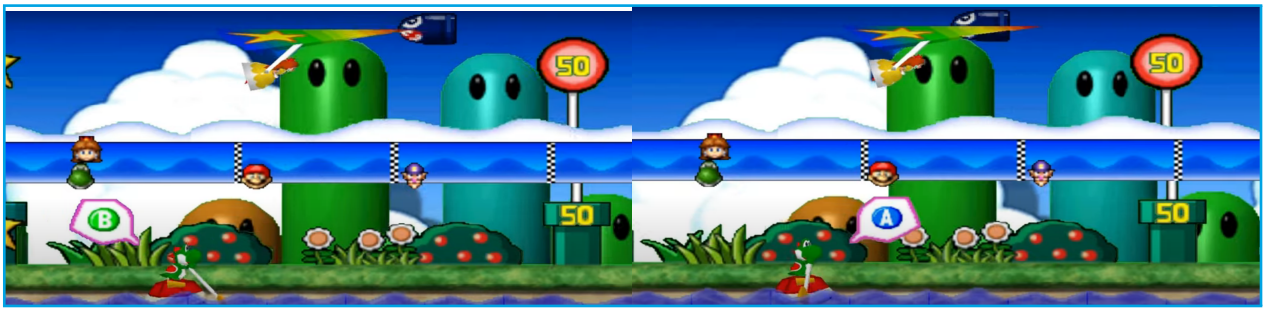


Figure 14.10: Ridiculous Relay (Mario Party 3) by Nintendo64Movies.

Finger pressing (Score: 90, Rank: 4) is the most important ability as it is used to paddle the boat. As overall progress for the 3-player team depends on the speed of each segment, the AIBM player is incentivized to be as fast as possible. **Inhibition** and **selective attention** are the next most important abilities. They keep the player from messing up the sequence by pressing the same button repeatedly (which stops the player from moving and costs them time). The short sequence length and input display means that these secondary abilities are supported and so overall noticeable, but perhaps not as limiting (Score: 40, Rank: 2). As with Psychic Safari, the sequence display allows for **token change detection** (Score: 20, Rank: 1) and **object recognition** (Score: 10, Rank: 1) to support the attention abilities by reminding the player what the next input should be if they make a mistake. **Tactile perception** and **procedural memory** are also used, but are unimportant compared to the other abilities (Score: 10, Rank: 1).

Speed Skating [423] has players race each other around the Olympic rink by primarily alternating pressing the left (L) and right (R) shoulder buttons[8] on their Nintendo DS. The control scheme is displayed on the top DS screen above the player's character, with the button(s) being pressed highlighted in yellow (Fig. 14.11). In the race setting, the time limit is implicit based on the other players or CPUs speed. The game displays this through the lap time (bottom right corner of the top screen), rink mini-map to show player positions (bottom screen), and Olympic Record (i.e. best time — bottom screen).

Finger pressing (Score: 90, Rank: 4) is important to this challenge as speed is the focus of the game. **Inhibition** and **selective attention** (Scores: 45, Rank: 2) keeps the player in sequence and not just "smashing" buttons. Messing up the sequence means the player stops skating and thus takes longer to finish their lap and jeopardizes their rank. They also keep the player from getting distracted by the multi-screen visuals which could influence their behaviour. While this means attention is more taxed than previous examples, these abilities are still supported by **token change detection** (Score: 25, Rank: 1) and

[8]The controls change slightly as players round corners. Then they have to hold the L button and mash the R button.

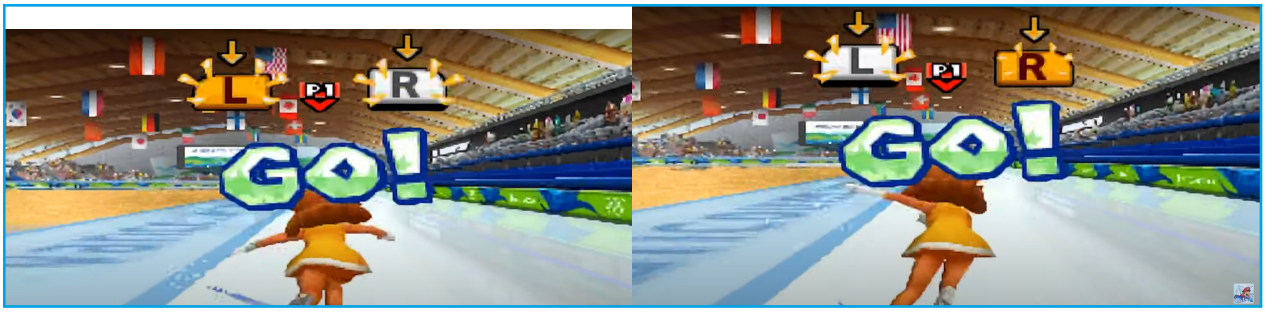


Figure 14.11: 500m Speed Skating (Mario and Sonic and the Olympic Winter Games 2010) by NintenU.

object recognition (Score: 10, Rank: 1) because of the visual display of the sequence and its current state. Token change detection also keeps us aware when the inputs change for going around a corner, or skating on the straight track, making it slightly more necessary in this example. **Tactile perception** and **procedural memory** are used as before (Score: 10, Rank: 1). Procedural memory may help with the layout and expectations of the game (i.e. knowing the change from straight to curve track). However, all of these low rank abilities are less important because of the redundancy measures for them built into the design through visual feedback.

Colossus of Rhodes [412] is a boss fight with many different phases and sub-challenges. One challenge at the end of the fight's second phase tasks the player with escaping from Colossus' hand by alternately mashing the L1 and R1 buttons (Fig. 14.12). The game displays the buttons that need to be mashed on the bottom left corner of the screen. The button that is currently pressed appears depressed on the icon, while the button that needs to be pressed appears unpressed. The game has an implicit time limit of approximately ten seconds; if the player does not input anything in that ten second period the Colossus will crush the player leading to an immediate game over. The aim of the AIBM segment is to complete it as fast as possible, with average players able to escape in three seconds

Figure 14.12: Colossus of Rhodes boss fight (God of War 2) by Boss Fight Database.

Finger pressing (Score: 90, Rank: 4) is the most important ability as quickly being able to press the buttons in sequence is the only way to succeed. **Inhibition** and **selective**

attention (Scores: 40, Rank: 2) keep us on sequence, and help us refrain from spamming buttons in a way that leads to mistakes. These abilities are important because of the extreme loss conditions. At this point in the larger boss fight, the player has finished an entire phase as well as a whole environment/dungeon to get back to this fight. A game over screen at this point takes you back to the beginning of this phase of the fight. The attention abilities are supported by **token change detection** (Score: 20, Rank: 1) through observing which button in the sequence is next. **Object recognition**, **tactile perception**, and **procedural memory** are also used (Scores: 10, Rank: 1) to: recognize the buttons to press, understand the buttons' state (pressed/not pressed), and remember the layout of the controllers to navigate the buttons easily. Procedural memory is slightly more used here because of the Playstation 2 controller having two sets of shoulder buttons.

14.3 Multiple Input Button Mashing (MIBM)

MIBM are motor focused challenges where players are "pushing multiple buttons simultaneously, repeatedly, and rapidly" [441]. Like SIBM, their main mechanics are button pressing and short time limits. The difference between the two categories is the increased load of having to press multiple buttons at the same time, versus a single button repeatedly. The main variables are number of buttons to be pressed, their location, and the time limit. In a multiplayer individual competitive context on a standard controller the limiting factor is finger pressing as it determines how many presses can be made in the time limit. Fig. 14.13 summarizes our initial competency profile.

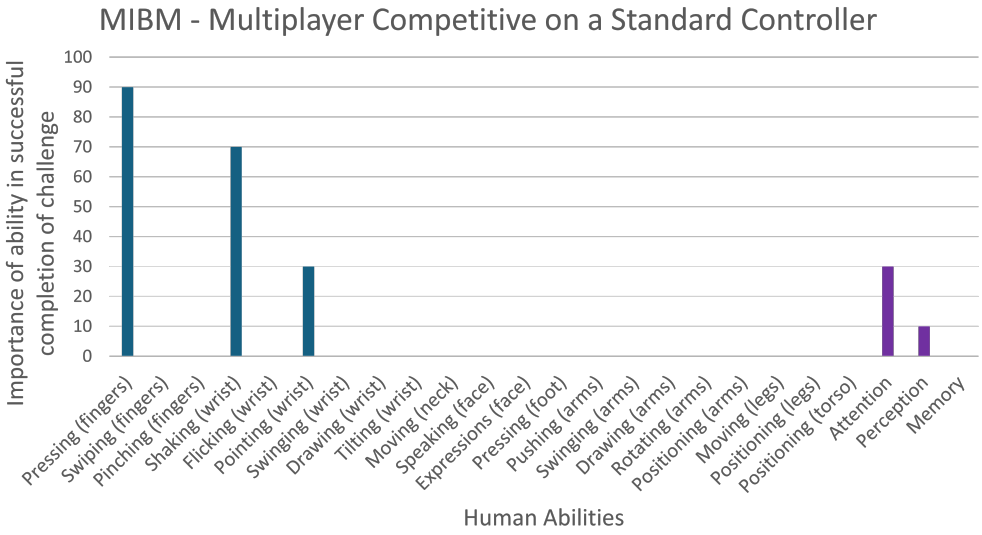


Figure 14.13: Challenge Description for Multiple Input Button Mashing as played on a standard controller from [441]

The original gameplay instances used to create this competency profile are:

- Mecha-Marathon, Mario Party 2 [194], Mario Party Superstars [323]; and,
- Chin Up Champ, Wii Party [348].

In revisiting the gameplay instances this information was drawn from, we arrive at a new competency profile which incorporates more specific cognitive abilities:

1. Finger pressing (Score: 90, Rank: Limiting),
2. Selective attention (Score: 35, Rank: Noticeable),
3. Inhibition (Score: 35, Rank: Noticeable),
4. Tactile perception (Score: 20, Rank: Used),
5. Object recognition (Score: 15, Rank: Used), and
6. Procedural memory (Score: 10, Rank: Used)

We reduce this list to the measurable abilities, presenting a hypothesized competency profile in Fig. 14.14.

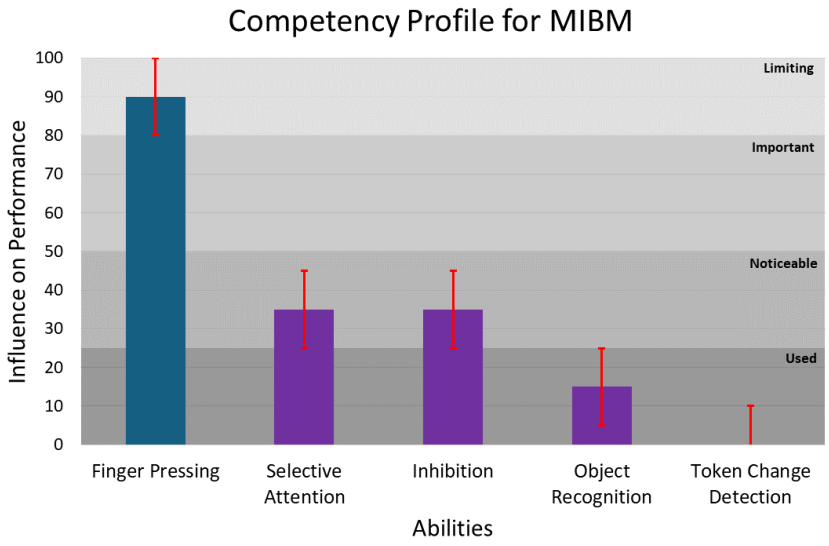


Figure 14.14: Hypothesized MIBM Competency Profile. Motor abilities in Blue, Cognitive abilities in Purple.

14.3.1 Gameplay-Ability Descriptions

We consider Mecha-Marathon the prototypical example, and so favour its distribution in our synthesis of the challenge competency profile.

Mecha-Marathon [194, 323] has players winding up Mecha Fly-Guys (Mechas) to race them, with the winner being the Mecha that flies for the longest distance. In Mario Party 2, the max distance is unknown, but (through tool-assisted experimentation by the speedrunning community) determined to be just under 54 metres (in-game). For Mario Party Superstars, the max distance is 70 metres (in-game). In both games, players have 10 seconds to wind their Mecha. The time is displayed in the top-centre of the screen (Fig. 14.15). Players wind their Mecha by repeatedly and simultaneously pressing the A and B buttons. The distance the Mecha travels is directly related to the number of times the player can mash the

buttons during the time limit. The winding animation does not correlate with the number of button presses registered (i.e. you will press the button more times than the animation can play).

Figure 14.15: Mecha-Marathon (Mario Party 2) by NintendoMovies.

Finger pressing (Score: 90, Rank: 4) is still the most important ability for Mecha-Marathon, but it is supported by wrist shaking and wrist pointing. On reflection, wrist pointing and shaking are actually ***strategic choices*** on the players part and not an inherent ability related to the challenge. Players with different strategies can successfully complete the challenge without these abilities, and so we remove them from the competency profiles. **Selective attention** and **inhibition** stop impulsive, incorrect behaviours. Since nothing on screen indicates your progress and the only thing that matters is your button presses, a player could look down at their controller the whole time and still win the game. So these are noticeable but not very important (Scores: 35, Rank: 2). This makes **tactile perception** (Score: 20, Rank: 1) more important in this challenge because the coordinated button mashing is more complex (the game will not count a non-simultaneous press). This means the haptic feedback of knowing we are pressing both buttons at the same time is key to correctly playing this game. As well since the winding animation does not correlate with our progress, this feedback is the only understanding we have of our success before scoring. **Object recognition** and **procedural memory** are used (Scores: 10, Rank: 1) in the same way they are for SIBM and AIBM.

Chin Up Champ [348] is a 4 person competition, where the player who does the most chin-ups wins. The players have 10 seconds to do as many chin-ups as possible. Chin-ups are performed by simultaneously pressing the A and B buttons on the Wii Remote. Players can perform at most 99 chin-ups (approximately 10 presses a second). The timer is displayed in the top-centre of the screen (Fig. 14.16).

Finger pressing (Score: 90, Rank: 4) is still the most important ability. As with Mecha-Marathon, wrist pointing supports this by allowing for faster pressing through wrist stability but this is inherently a strategic choice and so removed from the competency profile. **Selective attention** and **inhibition** play a slightly greater role here as there are more distractions on screen. While elements like the counter help make progress more obvious to the observer, it is still fundamentally unimportant for the gameplay (Scores: 40, Rank: 2). **Tactile perception** (Score: 20, Rank: 1) becomes important to know whether our input registered so we do not need to look at the screen as an indicator of our progress. **Object**

Figure 14.16: Chin-Up Champ (Wii Party) by NintendoMovies.

recognition and **procedural memory** are not noticeably used (Scores: 10, Rank: 1).

14.4 Summary of Button Mashing Challenges

We now have three hypothetical challenge competency profiles (SIBM, AIBM, MIBM). Through their shape we can begin to consider the relationship between player abilities and challenge requirements. Moving forward we need to test these hypothetical profiles to see if our assessments of the gameplay are relatively accurate.

Take home points

From this chapter we learned the following meta-lessons:

- Our initial competency profiles from Soraine and Carette [441] reasonably capture motor abilities.
- Extending cognitive abilities for challenges shows some deeper consideration for *how* the games are played (e.g. strategies versus basic requirements).
- Some abilities in the competency profiles cannot be measured, so we use reduced competency profiles.
- Button mashing family of challenges seem fairly stable given our closer look.

Chapter 15

Developing Understanding of Competency Profiles

Recall our thesis poses three specific research questions:

RQ1: How are cognitive and motor abilities used to interact with various challenges?

RQ2: What are the effects of cognitive and motor overloading on the mechanical achievability of a challenge?

RQ3: How can designers use this knowledge?

With our player profiles (Ch. 12) and competency profiles (Ch. 14), we now have the ability to answer these questions.

We conduct an exploratory study on player performance in, and experience of, button mashing challenges. We divide the study into three conceptual sub-studies: the first aims to validate our competency profiles (RQ1, Ch. 16), the second explores ability loading through mechanical difficulty (RQ2, Ch. 17), and the third connects the mechanical experiences (MX) to broader player experiences (PX) and player perceptions (RQ3, Ch. 18). We use this chapter to explain the design of the sub-studies, the structure of the overall study, and other information relevant to all three sub-studies. The data analyses and discussions for each sub-study are presented in their own chapters.

We first explain our sub-study designs and any relevant background for our choices. We then describe the apparatus and meta-experimental procedure. We then explain specific data collection and general cleaning/analysis processes. We end with some general limitations of our study and how we account for them. There are no take-home points for this chapter.

15.1 Study Designs

We design each study independently, and then combine common elements into a singular procedural flow.

15.1.1 Validation study design (Study 1A)

We validate our competency profiles by showing a quantifiable relationship between our predicted challenge abilities and the player's performance.

We create a complex correlational study, where all participants play all the challenges tuned at the same level. We use *multiple linear regression* (MLR)[1] [168] to find the abilities that affect challenge performance. We regress participant challenge scores (dependent variables) on their abilities (independent variables) to create models for each challenge (Table 15.1).

Independent variables	Finger Pressing, Selective Attention, Inhibition, Object Recognition, Token Change Detection
Dependent variables	SIBM Score, AIBM Score, MIBM Score

Table 15.1: Independent variables (measured abilities) and Dependent variables (in-game performance) for our study.

15.1.2 Loading study design (Study 1B)

We explore the effect of loading on competency profiles by having participants play challenges tuned to different difficulty levels via mechanical difficulty.

We setup a series of single-factor multi-level experiments using a between-groups design. Our independent variable is game difficulty (i.e. a score modifier proxy for finger pressing load), and we set it to three levels (Easy: 2, Control: 1, Hard: 0.5). We assign participants to conditions via block randomization to ensure that each group is the same size. We compare group gender distributions, ages, and abilities to identify whether these are confounding variables. Participants play each game once, tuned to their difficulty condition. We counterbalance the game orders (6 unique sequences) to distribute any ordering effects across groups. For each condition, we regress the participant scores for each game on the abilities identified in the validation study to create competency profiles for each loading. We draw preliminary conclusions about the relationship between limiting ability load and performance by comparing these models to each other and the validated baselines.

15.1.3 Experience study design (Study 1C)

We use a convergent mixed-methods research design (*questionnaire variant* [100]) to study player perceptions and PX of the challenges at different difficulties. We *simultaneously* collect, analyse, and integrate quantitative and qualitative data to create a more robust understanding of a research question [101].

We use the Player Experience Inventory (PXI) to gather PX data (quantitative Likert data). We append open-ended questions to the PXI to gather participant opinions on each game's difficulty (qualitative textual data). We analyse the Likert data via One-Way

[1] A background on MLR can be found in App. F.

ANOVAs and post-hoc Tukey HSD tests, and use thematic analysis for exploring the textual data. We integrate the data through a side-by-side comparison; we present the results independently, and then compare them in a meta-discussion [101].

15.2 Apparatus

All sub-studies use the same hardware. Participants only use MGBatt in the first sub-study (Study A: Validation), but its data is used in the analysis for both Study A (Validation) and Study B (Loading). Different configurations of the Custom Button Mashing Games are used in Study A (Validation), and Study B (Loading). The data from both iterations of the Custom Button Mashing Games is used as contextual information in Study C (Experience). The data collection tools are predominantly used in Study C (Experience).

15.2.1 Hardware

The testing stations are identical to the ones from Ch. 11 (see Fig. 11.2). Each station has:

- A Windows 10 desktop computer (i5-4670k 3.40 GHz processor, 16 Gb RAM, Nvidia Geforce GTX 780Ti graphics card);
- A single 30-inch monitor;
- A K120-TAA Logitech full size wired USB keyboard;
- A B100-TAA Logitech wired USB mouse; and,
- A pair of Beyer Dynamic DT 990 over the ear headphones.

Participants sit 26-inches away from the monitor to maintain ergonomics.

A Quick Aside...

For these studies, we must define what it means to "push a button". Depending on the keyboard type, a "press" could mean anything from a light touch or a full hard slam. We consider one button press to require a complete depress of the button. Therefore we use membrane keyboards which require a complete depress to register a "press" signal. This forces participants to be consistent with each other about what it means to "press a button", and therefore what "counts" towards their score.

15.2.2 Software

This experiment uses two custom built softwares: the Mini-game Ability Battery, and three custom button mashing mini-games.

Mini-game Ability Battery (MGBatt)

MGBatt (details in App. A.2) consists of 5 mini-games that test finger pressing (FP), selective attention, inhibition, object recognition, and token change detection, as mapped in Table 15.2. Participants play each mini-game three times, at increasing difficulty levels. Table

11.1 from the previous study describes the configurations for the difficulty conditions. The mini-game order is randomized. We measure a player's ability levels in the units from Table 15.2. **We use the player's performance measures in the regression modeling**.

Game	Ability	Measurement
Digger	Finger Pressing	(P/S) × 10
Looking	Selective Attention* /Inhibition	C/S
Recipe	Selective Attention /Inhibition*	C/S
Cake	Object Recognition	C/T
Stage	Token Change Detection	C/S
Legend:	Total presses of correct input across conditions (P), total response time across all conditions in seconds (S), Total correct responses across conditions (C), Total trials across conditions (T)	
Note:	In cases where a mini-game measures multiple abilities, we have starred (*) the ability it is named in the data.	

Table 15.2: Ability measurement games summarized.

Custom button mashing games:

(a) Fire Starter (b) Fly Away (c) Potion Master

Figure 15.1: Three custom button mashing games designed to test the button mashing competency profiles.

We create three mini-games, representing each type of button mashing challenge[2]: *Fire Starter* (SIBM, Fig. 15.1a), *Fly Away* (MIBM, Fig. 15.1b), and *Potion Master*(AIBM, Fig. 15.1c). In these mini-games the player is a young witch practicing her magic skills. Each game operates abstractly in the same way: participants must press the correct inputs as many times as possible within the given time limit. Each game calculates the player's score as the number of correct inputs multiplied by a score modifier (to represent the "difficulty level"). The game then ranks players by comparing their performance to a target/goal value (see Table 15.3). The ranks are irrelevant for validating or exploring competency profiles, but they incentivize players to put in reasonable effort and promote "game" feel. **We use the player's score to represent their challenge performance in regression modeling.**

[2]All challenges are developed in Unity 2022.3.0f1 using the C# scripting language. The source code for the games, along with an executable, are available on GitHub.

Rank	Player final score	Effect Meaning
F	$< 0.5{*}$Goal	Really bad
D	$0.5{*}$Goal $\geq$ Score $< 0.75{*}$Goal	Bad
C	$0.75{*}$Goal $\geq$ Score $<$ Goal	Low Average
B	Goal $\geq$ Score $< 1.25{*}$Goal	High Average
A	$1.25{*}$Goal $\geq$ Score $< 1.5{*}$Goal	Good
S	$\geq 1.5{*}$Goal	Really good

Table 15.3: List of ranks and their interpretation based on player final score relative to the goal number.

The mini-games are configurable; details about their design decisions and variable parameters are covered in Appendix D. We present the configurations for this study in Table 15.4. We use the average FP rate in the previous study (Ch. 11) and play testing done on the games to set the goal values. We adjust load on FP by changing what "counts" as correct input via the score modifiers. "Easy" underloads FP (reducing load by factor 2), and "Hard" overloads FP (increasing load by factor of 2).

	Buttons		Time limit	Goal	Score Modifiers		
	1	2			Base	Easy	Hard
Fire Starter	→	–	10s	77	1	2	0.5
Potion Master	→	←	10s	50[3]	1	2	0.5
Fly Away	→	←	10s	77	1	2	0.5

Table 15.4: Configuration parameters for all button mashing challenges.

15.2.3 Data Collection Tools

Pre-study Survey (App. E.2): We ask 15 questions broken into demographics, gamer identity, and gaming history sections. For demographics, we collect the participant's self-identified gender, age, education, self-identified disability status, and if applicable any impacts on abilities and assistive devices used. We focus on gender instead of sex as we consider how self-identity impacts a participant's relationship with gamer identities and gaming generally. We use "self-identified" language for gender and ability questions to reflect that participants may have barriers to accessing formal labels and diagnoses, or may not be comfortable reporting and presenting this information in a lab setting. Gamer identity is either Yes, No, or a write-in "Other" response. Through the "other" option, we capture more complex relationships to the gamer identity, such as being unsure if they are a gamer or feeling like they used to be a gamer. For gaming history we gather how long they have been playing games, their frequency and length of their gaming sessions, the devices they commonly use to play games, their familiarity with a keyboard and mouse control scheme, and the types of games they currently or previously used to play. We leave the types of games question (Q12) as a

[3]This is the number of correct sequences of Button 1 - Button 2 presses.

write-in allowing for participants to either list genres, or games that we can then interpret during data processing.

Player Experience Inventory (PXI) [4]: We ask 24 questions, covering all constructs except Autonomy and Curiosity (see Table 15.5). We recognize the PXI is designed for gathering experience around more complex games and longer gaming sessions. However, there are no validated inventories for experiences with mini-games specifically, so we aim to use PXI but are aware these results will be affected by the lack of nuance.

We left out Autonomy and Curiosity due to the fact we were testing mini-games. Given the simplicity of the mechanics, short gameplay, and lack of larger game, questions like "I felt a sense of freedom about how I wanted to play this game." (Autonomy), and "I felt eager to discover how the game continued" (Curiosity) did not seem applicable to the context of our study. The other constructs have a stronger theoretical overlap with MX and are possible to inspire in a short session like a mini-game, so we include them in case there were any interesting results.

Open-ended Questions: We append four, write-in responses questions to the end of the PXI. The first three ask participants to self-reflect on their experience for each game and communicate the elements they found particularly difficult (e.g. "What did you feel was the hardest part of the fire-starting game?" to ask about SIBM). The final question explicitly asks participants "Are there any comments you would like to leave the researchers about your experience with the games?" This question gives participants an open forum to note down thoughts that may contextualize their responses/experience, and an outlet to describe any feelings not related to the specific button mashing challenges.

Observational Notes: We capture participant behaviour during the custom mini-games in informal notes. We particularly focus on their strategy choice (one-hand vs. two-hand strategy, particular holds for the keyboard, where on their body the movement was happening), seeming emotional response to the baseline and experimental conditions, and any notable comments or behaviours they made during gameplay or directly after.

Meaning	
Mean1	Playing the game was meaningful to me.
Mean2	The game felt relevant to me.
Mean3	Playing this game was valuable to me.
Mastery	
Mast1	I felt I was good at playing this game.
Mast2	I felt capable while playing this game.
Mast3	I felt a sense of mastery playing this game.
Immersion	
Imm1	I was no longer aware of my surroundings while I was playing.
Imm2	I was immersed in the game.
Imm3	I was fully focused on the game.
Progress Feedback	
Prog1	The game informed me of my progress in the game.
Prog2	I could easily assess how I was performing in the game.
Prog3	The game gave clear feedback on my progress towards the goals.
Audiovisual Appeal	
AV1	I enjoyed the way the game was styled.
AV2	I liked the look and feel of the game.
AV3	I appreciated the aesthetics of the game.
Challenge	
Cha1	The game was not too easy and not too hard to play.
Cha2	The game was challenging but not too challenging.
Cha3	The challenges in the game were at the right level of difficulty for me.
Ease of Control	
Con1	It was easy to know how to perform actions in the game.
Con2	The actions to control the game were clear to me.
Con3	I thought the game was easy to control.
Goals and Rules	
Goal1	I grasped the overall goal of the game.
Goal2	The goals of the game were clear to me.
Goal3	I understood the objectives of the game.

Table 15.5: Subset of Player Experience Inventory questions for our study.

15.3 Procedure

Fig. 15.2 shows our meta-procedure. We indicate the different studies using labeled, coloured blocks.

Pre-study: We first collect the participant's written informed consent. We lead the participant to a testing stations where they complete our pre-study survey (App. E.2). If both testing stations are in use, the survey is available on a laptop at the waiting area (couches in Fig. 11.1).

15.3.1 Study 1A: Validation Study

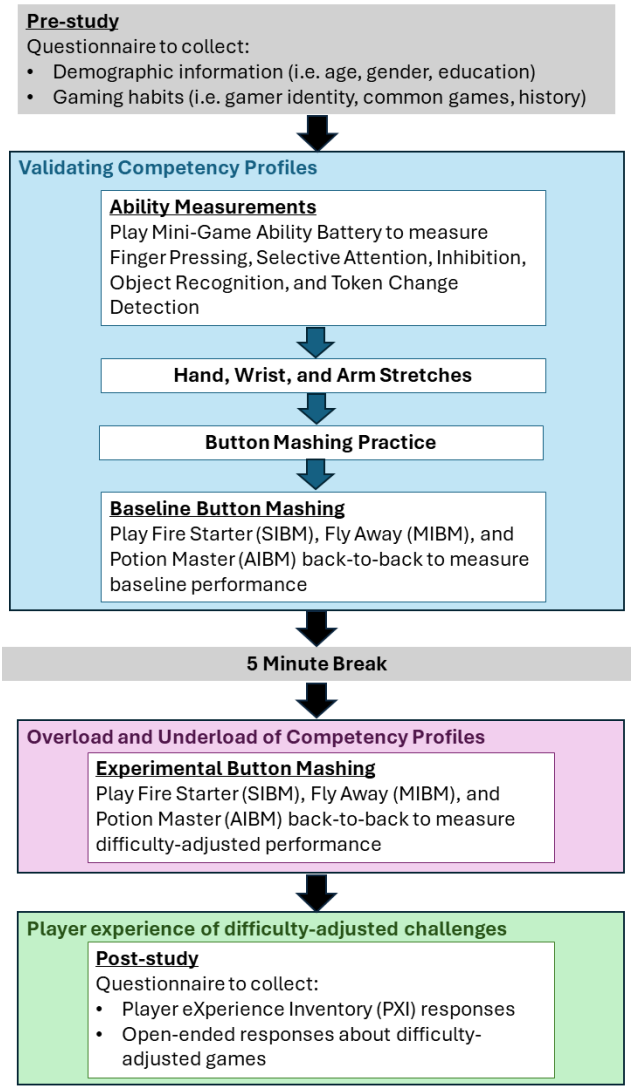


Figure 15.2: Overall study procedure. Study 1A (Validation) is the blue block, Study 1B (Loading) is the pink block, and Study 1C (Experience) is the green block.

Ability Measurements: After completing the survey, we start participants on the MGBatt. We inform them of the approximate time it will take, what to expect of the games, and how to indicate if there is a technical issue. We allow the participant to ask us any questions if the tasks are unclear. Once the participant completes their MGBatt, we lead them in a series of hand and wrist stretches to warm up for button mashing and give their eyes some rest from the screen.

Baseline Button Mashing: We start the custom mini-games. We allow participants to practice each mini-games "until they feel ready to do the baseline measurement". We advise them that practice games use the same mechanics as the baseline, but are slightly easier than the real measure[4] and do not provide a clear rank. We explain this is so the participant does not fatigue themselves in the practice. Participants may ask us any clarifying questions about how to play, the game mechanics, or how to interpret their feedback. Once the participant feels they understand the games and how to play, they start the real measure. Participants play the button mashing games back-to-back with the parameters from Table 15.4. After completing the games, the participant takes a five-minute break.

15.3.2 Study 1B: Loading

Experimental Condition: We assign participants a letter-group to determine their condition (A: Easy, B: Control, C: Hard). We do not give participants practice time in the adjusted condition, as they already understand the mechanics from the baseline. Participants play the games in a random order, which could be different from their baseline order.

[4]The practice games use the default values for goals and modifiers as outlined in App. D.

15.3.3 Study 1C: Experience

Post-Study: Participants immediately complete the post-study survey of: 24 PXI questions, 3 additional Likert-items explicitly asking whether players enjoyed the game, 3 write-in questions about what they found difficult about each game, and an open-ended question that allowed them to give us any feedback. On average participants complete the survey in six minutes and twenty seconds. We then compensate participants $20, pro-rated at $10 per half hour, for participating in this study[5].

15.4 Data Collection and Analysis

75 participants complete our experimental session. Table 15.6 summarizes the data we collect and its formats. The anonymized cleaned data for each study, custom Python files, Matlab files and Matlab outputs are available at Sasha's GitHub for review/use.

Data	Captured via...	Automatic/Manual	Output
Ability measures	Ability Battery	Automatic	JSON
Game performance	Mini-games	Automatic	JSON
Pre-study Survey	Microsoft Forms	Manual	Excel
Post-study Survey	Microsoft Forms	Manual	Excel
Observational Notes	Sasha's notes	Manual	Excel

Table 15.6: Data collected during the experimental session.

Processing Data: We use custom Python scripts to process and convert the JSON files into Excel workbooks. We merge all the Excel workbooks into a single dataset containing each participant's ability data, experimental condition, pre-study data, post-study data, and strategy (one vs. two handed) from observational notes.

Cleaning Data: For each study we manually clean the dataset based on the study's specific needs. This involves removing incomplete or irrelevant participant data, not removing outliers. Outliers are not removed because individual differences between participants could make significant deviation from the mean plausible and reasonable.

Analysing Data: We analyse participant demographic data in Excel. We use the Matlab 2024a [283] Statistics and Machine Learning Toolbox [282] to calculate descriptive statistics and perform regressions. We use IBM SPSS v29 [207] to calculate One-way ANOVAs and Tukey HSD Tests for PXI responses.

15.5 Common Limitations

We summarize some meta-elements that limit our study (Table 15.7).

[5] All participants completed their sessions so no pro-rated payments were made.

Potential limit	**Study it affects**
Participant homogeneity	1, 2
Human variability	All
Fatigue	All
Gaming experience	All
Modulating effort	1, 2
Ability Battery uncertainty	1,2
Difficulty scaling uncertainty	2,3
Small sample size	1, 2
Validating study (1), Loading study (2), Perceived experience (3)	

Table 15.7: Potential limitations and the studies they affect.

Participant Homogeneity. We recruit participants age 18+ around McMaster Campus. While we did not explicitly recruit gamers, our study about games attracted more gamers than not. **This makes our sample more homogeneous than expected, and so our data may be more peaked than normal, or skewed to reflect this**.

Human variability. Humans are not precise performers, and so across multiple measurements we expect a degree of **variability caused by individual differences and environmental factors**. As well, we cannot control the participants lives around the study; for example, one participant came to the session directly from working out at the gym, which could affect their performance. We account for some of this variability by taking baseline and experimental measures in the same session to reduce differences between environmental factors.

Fatigue. Participant could experience fatigue from the length of the experimental session, and the repetitive nature of the tasks. We account for some of this by:

- incorporating stretches in between the Ability Battery and Baseline measure;
- randomizing the game orders inside the MGBatt and experimental games; and,
- including a 5 minute break between the Baseline and Experimental condition.

Gaming experience. Button mashing as a challenge (as opposed to a strategy) was common during the early 2000s as part of quick-time events and other mini-games, however it is not as common in popular gameplay post-2010. Given the average age of our participants, those who have been playing for 10 years were gaming in this post-button mashing challenge era. We see this reflected in the commonly cited genres and games (e.g. Minecraft, Valorant) which prioritize cognitive-focused gameplay challenges. **Experienced (i.e. *literate*) players (e.g. gamers) develop a sort-of "professional vision" (effectively being able to ignore other stimuli irrelevant to the task [164]) for games [61, 218] which can improve their performance.** For button mashing challenges, literate players may ignore on-screen feedback and cues, choosing to focus exclusively on the input device as the only relevant component in task success. Illiterate participants, like those only familiar with button-mashing-as-a-strategy, are more likely to focus on visual feedback since that reflects

what they think is important to the task. **This experience also affects the participants' strategies, as they believe different grips and holds (e.g. one handed or two handed), afford optimal performance**. While minor differences may create outliers in the data, sufficient splits based on strategies would create bimodal data. Since we cannot control any groups that form because of strategy we may not have enough participants in each group to run a reasonable regression. **We try to offset the personal experience, and strategy impact by allowing for open-ended practice before attempting the baseline challenge**. Through this practice we anticipate players can try out different strategies and become familiar with the mechanics such that they are more consistent in approach between baseline and experimental condition.

Modulating Effort. Participants modulate their effort in a game based on their expectations about its difficulty and their performance. It is possible that participants can deduce their difficulty level (i.e. condition) during the experiment by comparing their baseline and experimental ranks. This **could impact whether they hold back in their effort or double-down**. Future work could try and account for this by dynamically adapting the difficulty based on the participant's measured ability score. However, this could actually exacerbate effort modulation. Consider a participant in the Easy condition who realises the game is trivial and so reduces their effort. Our current fixed difficulty conditions mean that even inside the "Easy" condition some participants may perceive it as significantly easier or harder due to their personal ability. So while our imaginary participant reduces their effort, another group member may increase their effort because they find it more challenging. If the game was dynamically adjusted, it is possible that all Easy condition players will respond the same way and reduce their efforts.

Ability Battery Uncertainty. We measure the participants abilities through MGBatt's mini-games. While the measurements seem reasonable, inherently they are trying to approximate an ability level through a proxy mini-game. **There is an inherent level of uncertainty in these measurements that we accept as part of the measurement process**.

Uncertainty from difficulty scaling. We change difficulty by scaling the participant's number of correct button presses (i.e. score modifier). **This also scales any latent noise from human variability and participant gaming experience as well**. We expect the uncertainty between conditions is relatively scaled with each other, and can check this against the prediction intervals for each condition's regression models.

Small Sample Size. Our small sample size could **affect the normality of our data, increase the effect of noise, reduce the power of our regressions, mislead us about correlations between predictors and the dependent variable, and make it difficult to meet the regression assumptions**. **This is less likely to affect the validation study** as we have more than 10 observations per predictor, which is a commonly cited (though frequently debated) rule of thumb [368]. It is **more problematic for the loading study** since each condition has fewer than 30 data points and multiple predictors.

Considering this is an exploratory study, we believe that the power is less important as seeing general trends and changes. We also assess the underlying regression assumptions for each resulting regression to discuss their appropriateness on a case-by-case basis.

15.6 Heading into studies...

We now head into the study results for RQ1 (Ch. 16), RQ2 (Ch. 17) and RQ3 (Ch. 18). Each chapter introduces its study and has a short background on literature relevant to the study design and procedures. Since the data subsets are different for each study, each chapter has its own participants section to outline whose data is used.

Chapter 16

Validating Challenge Competency Profiles (Study 1A)

We conduct an exploratory study to validate our competency profiles by comparing them against regression models for player scores using player abilities. This gives us a sense of how successful our close reading method is at identifying the general competency profile of a challenge, and provides us with answers to RQ1.

We begin this chapter explaining why multiple linear regression is appropriate, and elements of the process we must consider for our analysis to be reasonable. We explicitly lay out our hypotheses regarding our competency profiles and the regression models. We then summarize our study design, and participant information. Having laid this groundwork we present our data and resulting models. We discuss whether the models meet regression assumptions, and factors that impact their performance. From there we compare our regression models to our competency profiles and discuss the ways in which they align and differ.

16.1 Background: Multiple Regression

For button mashing challenges, we think that a player's score, S, can be expressed as a linear combination of some subset of Finger Pressing (FP), Selective Attention (SelAtt), Inhibition (Inhib), Object Recognition (OR), and Token Change Detection (TCD), such that:

$$\mathrm{S} \sim 1 + \mathrm{B}_{FP} \times \mathrm{FP} + \mathrm{B}_{SelAtt} \times \mathrm{SelAtt} + \mathrm{B}_{Inhib} \times \mathrm{Inhib} + \mathrm{B}_{OR} \times \mathrm{OR} + \mathrm{B}_{TCD} \times \mathrm{TCD}$$

where $\mathrm{B}_{Ability}$ is a scalar coefficient associated with the ability variable, 1 is a placeholder for a constant valued intercept, and ability units are game dependent (summary in Table 16.1). A linear model seems reasonable given the history of its use in performance-based games research (e.g. 116, 132), and our lack of theoretical reasons to look at a non-linear model.

We construct the models through *multiple linear regression (MLR)*, using a *stepwise* process to find an optimal subset of predictors [168][1]. While criticised, stepwise regressions are considered acceptable for small, exploratory studies like ours, where little-to-no theory exists or where researchers are looking for insight into a hypothesized theory [185]. We

[1]Details about MLR and Stepwise regression can be found in App. F.

start with an empty model (i.e. just a constant), and add/remove predictors at every step to optimize for a small sum of squared error (SSE)[2] as determined by a significant F-Test at $p < 0.05$ between model iterations. We select SSE over R^2 because R^2 will generally increase with more predictors, thus biasing the resulting model. The resulting reduced model is checked against the assumptions and diagnostics for MLR to determine if it is "reasonable" (details in Sec. F.0.2).

16.1.1 Assumptions of Regression Modeling

A reasonable MLR model needs to meet the following assumptions:

1. Each independent variable (IV) correlates to the dependent variable (DV);
2. The IVs are not collinear with each other;
3. The errors follow a normal distribution;
4. The errors are independent; and,
5. The errors are homoscedastic.

Correlation. Only IVs that correlate to the DV should be in the final model. Given the exploratory nature of this work, we do not set a threshold for sufficient correlation. We instead focus on significance at $p \leq 0.05$, as a stricter α did not seem necessary.

Collinearity. Highly correlated IVs should not be in the same model. We use a pairwise correlation matrix and variance inflation factors (VIF) to check for collinearity. So long as the IVs in the final model are not unreasonably correlated ($r < 0.5$) and the VIFs are less than 3, we consider the collinearity to be non-problematic. We choose these looser constraints because human abilities overlap and so we do not have perfectly orthogonal ability measures (see Ch. 11).

Residuals (i.e. errors) are normally distributed. We visually check residual normality through histograms, probability (P-P) plots, and quantile-quantile (Q-Q) plots. While we generally expect a normal distribution of errors, skew in the raw data due to participant demographics (age) and variable gaming history could cause tailing in the normality plots. So long as the normality plots do not deviate in *unexpected* ways, we say the normality assumption is met.

Residuals are Independent and Homoscedastic. We visually check these through a residuals versus fitted values plot. For these plots, we are looking for the data points to be randomly distributed with no specific patterns (*independent*) and for the data points to be relatively symmetrical around the 0-line (*homoscedastic*). We expect that the same skew from particularly "skilled" participants will cause some extreme outliers, but so long as no obvious pattern exists we believe these are acceptable.

[2]The SSE is a measure of the model's estimation power based on the difference between the real values and model's predicted values. A smaller SSE means the regression fits the data better, and gets us closer to a more precise model.

16.2 Hypotheses

Hypotheses:

Our competency profiles are "good enough" if...

H1: The reduced model uses all abilities which are listed above the threshold of "Noticeably used, but not important (Rank 2)";

H2: The abilities in the reduced model preserve the general order of importance for the abilities.

H1 is easy to evaluate by comparing the terms in the regression model to the abilities in the competency profile. We evaluate H2 by proxy through comparing the model's standardized coefficients, and the order abilities are added to the model via the stepwise process (details in 16.1). Our actual models use non-standardized coefficients to preserve the ability units, making them easier to discuss in a meaningful way. However, we calculate the standardized coefficients for the purpose of H2.

16.3 Study Design

We conduct a complex correlational study with one-group design; therefore every participant will have one set of results from the ability measures, and one score from each of the three button mashing games. Details about the study design, procedures, and apparatus are in Ch. 15.

16.3.1 Participants

We remove seven participants due to technical errors in the measurement process dropping some of their data (N= 68; 42 men, 22 women, and 4 queer[3])[4]. Participants are between 18 to 37 years-old ($\overline{\text{Age}}$: 21.9 years, σ: 4.2), and the majority report having some undergraduate education (74%). 63% self-identify as gamers (32 men, 8 women, 3 queer), and the majority (63%) report playing games for over 10 years[5]. They play frequently during the week (43 % play 5+ times a week; 34% play 2-4 times a week) commonly for 1-3 hours per session (50%)[6]. They predominantly play on their computers (84%) and phones (59%), and seem to focus on Action (38%) games or Action-Adventure (17%) — the most commonly named ones being Minecraft (16%), FIFA (16%) and Valorant (15%)[7].

[3]Queer is used as shorthand for participants who identify as non-binary, trans, queer, or unidentified.

[4]Gender was recorded as opposed to sex in order to see whether gender identity affected player experience in secondary study.

[5]The second largest group was 5 to 9 years at 21%

[6]The second most commonly reported session is 30 minutes to 1 hour at 25%

[7]The next three by name are League of Legends, Fortnite, and Call of Duty. "Battle games" are listed as a genre by players, which describes multiplayer action games centred around combat.

16.4 Study Results

All statistical processes are done via Matlab 2024a[283] using the Statistics and Machine Learning Toolbox[282] (see documentation [281]), except for VIFs which we calculate using Daniel Vasilaky's vif function [504].

16.4.1 Descriptive Statistics

Table 16.1 summarizes descriptive statistics and histograms for the participant's abilities (IVs). With the exception of OR[8], the ability measures seem reasonably normally distributed on visual inspection of the histogram — as we would expect in a general population. The data shows some "peaking", and SelAtt and Inhib skew right. Both could be due to our somewhat homogeneous sample (i.e. young, highly-educated, able-bodied, gamers) performing well and relatively similarly.

Looking at the game scores (DVs) descriptive statistics (Table 16.2) we see the distributions are reasonably normal on visual inspection of the histogram. SIBM has more of a left skew and is heavily tailed, but this could be an artifact of our sample having more practice and being at peak ability age. The central tendencies are similar between SIBM, MIBM, and Finger Pressing, indicating that the performance in those games may be similar.

16.4.2 Results of Multiple Regression Modeling

We run stepwise MLR between each button mashing game's score (SIBM, AIBM, MIBM) and the measured abilities (FP, SelAtt, Inhib, OR, and TCD) resulting in three statistically significant models:

- SIBM Score $= 26.70 + 0.68 \times \text{FP}$
- AIBM Score $= 20.75 + 0.28 \times \text{FP} + 11.38 \times \text{SelAtt}$[11]
- MIBM Score $= 28.38 + 0.55 \times \text{FP}$

Our models use unstandardized coefficients so they retain their units making discussion of the coefficients more meaningful. **The intercept of our models does not depict the score when a player's ability is 0**. Our measures are a proxy for the underlying ability, and so the concept of a 0 score is fuzzy. Consider that even a baby would have some amount of FP ability (hence being able to flex fingers), and so a "0" measure is likely impossible. If it was possible, we do not know if the increase in score between 0 and

[8] From mini-game battery, the game may be too easy even at this tuned level and so the measures have almost everyone getting perfect.

[9] Recall this represents the score from the Looking mini-game. Looking measures both selective attention and inhibition. It was labeled selective attention for ease of data exploration.

[10] Recall this represents the score from the Recipe mini-game. Recipe measures both selective attention and inhibition. It was labeled inhibition for ease of data exploration.

[11] A reminder that Selective Attention refers to the score generated from the Looking mini-game in the Ability Battery. Looking measures both Selective Attention and Inhibition since they are highly coupled abilities, and this value should be thought of as representing both abilities.

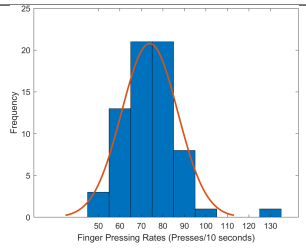
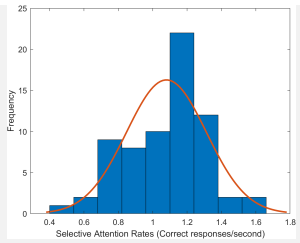
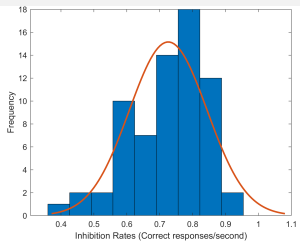
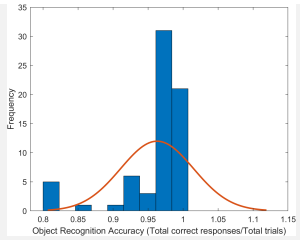
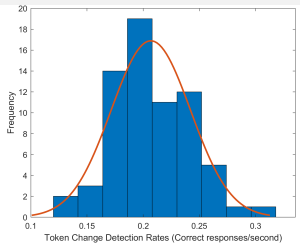
Histogram	Information	Take-away
	Ability: Finger Pressing Measure: Presses / 10 seconds Mean: 73.87 Std. Dev: 13.04 Median: 73.43	• Reasonably normal • Mean matches expectations • High performing outlier
	Ability: Selective Attention[9] Measure: Correct responses / second Mean: 1.08 Std. Dev: 0.23 Median: 1.14	• Reasonably normal
	Ability: Inhibition[10] Measure: Correct responses / second Mean: 0.73 Std. Dev: 0.12 Median: 0.74	• Skewed right; could be because of participant demographics • Still reasonably normal
	Ability: Object Recognition Measure: Accuracy (correct responses) Mean: 0.96 Std. Dev: 0.05 Median: 0.98	• Skewed right • Not normal
	Ability: Token Change Detection Measure: Correct responses / second Mean: 0.21 Std. Dev: 0.04 Median: 0.20	• Reasonably normal

Table 16.1: Summarized information about the independent variables (ability measures).

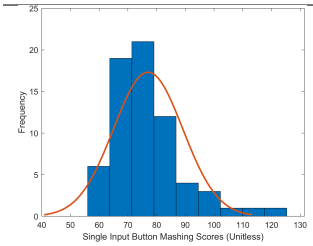
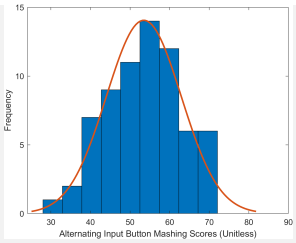
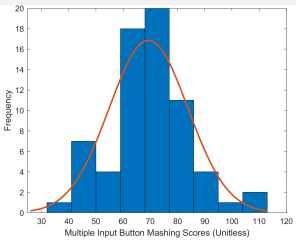
Histogram	Information	Take-away
	Game: SIBM Mean: 77.06 Std. Dev: 12.06 Median: 75	• Left skew • Heavy tail
	Game: AIBM Mean: 53.47 Std. Dev: 9.46 Median: 54	• Reasonably normal • Missing right tail
	Game: MIBM Mean: 69.35 Std. Dev: 14.48 Median: 69.5	• Reasonably normal • Peaking at the centre

Table 16.2: Summarized descriptive statistics for dependent variables.

1 is linear, or if scores react linearly below the data points we see. **The intercept is an artifact of the regression that captures some of the noise and error in our data, but nothing more**.

Regression Modeling:: SIBM Score $= 26.70 + 0.68 \times$ **FP**

Variable	Coefficient	CI (95%)	Std. Error	t-stat
Finger Pressing	0.68****	[0.53, 0.84]	0.08	8.87
Intercept	26.70****	[15.19, 38.20]	5.76	4.63
	$R^2 = 0.54$	$F_{(1,66)} = 78.7$****		
Significance: not significant ( ), $p < 0.05$(*), $p < 0.01$(**), $p < 0.005$(***),$p < 0.001$(****)				

Table 16.3: Stepwise regression model for SIBM.

The SIBM regression produces a significant model, $F_{(1,66)} = 78.7$, $p < 0.0001$, R^2=0.54 (details in Table 16.3). FP is a significant predictor ($t(66) = 8.87$, $p < 0.0001$) with a strong positive effect on Score. Each additional press per 10 seconds increases the player's score by

0.68 units. Plotting the regression line against the data we see it reasonably approximates a linear looking trend (Fig. 16.1).

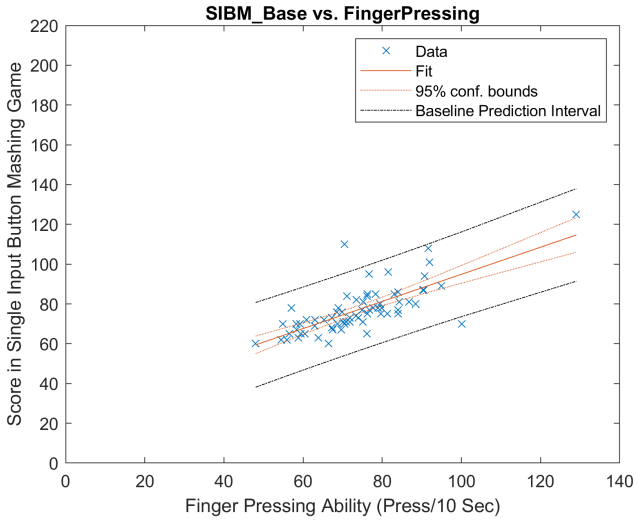


Figure 16.1: SIBM regression model plotted with the raw data.

Regression Modeling:: AIBM Score $= 20.75 + 0.28 \times$ **FP** $+ 11.38 \times$ **SelAtt**

Variable	**Coefficient**	**CI (95%)**	**Std. Error**	**t-stat**
Finger Pressing	0.28****	[0.12, 0.43]	0.08	3.52
Selective Attention	11.38*	[2.59, 20.17]	4.40	2.59
Intercept	20.75***	[7.33, 34.16]	6.72	3.09
	$R^2 = 0.27$	$F_{(2,65)} = 12.1$****	*Durbin-Watson = 1.98*	
Significance: not significant ( ), $p < 0.05$(*), $p < 0.01$(**), $p < 0.005$(***),$p < 0.001$(****)				

Table 16.4: Stepwise regression model for AIBM.

The AIBM regression produces a significant model, $F_{(2,65)} = 12.1$, $p < 0.0001$, R^2=0.27 (details in Table 16.4). FP ($t(65) = 3.52$, $p < 0.001$) and SelAtt ($t(65) = 2.59$, $p = 0.011955$) are significant predictors with positive effects on Score. For each additional press per 10 seconds, the player's score increases by 0.28 units; each additional correct response per second, increases the player's score by 11.38 units.

We plot the model-plane against the data (Fig. 16.2), and see that Score grows steeply in line with FP. Based on the SSE Criterion, FP seems to have a stronger influence on the Score ($F = 16.19$, $p < 0.001$), where SelAtt's effective change in the model was smaller ($F = 6.69$, $p = 0.011955$). The score's growth relative to each predictor is more visually apparent when looking at the slices of the plane at fixed predictor values (Fig. 16.3).

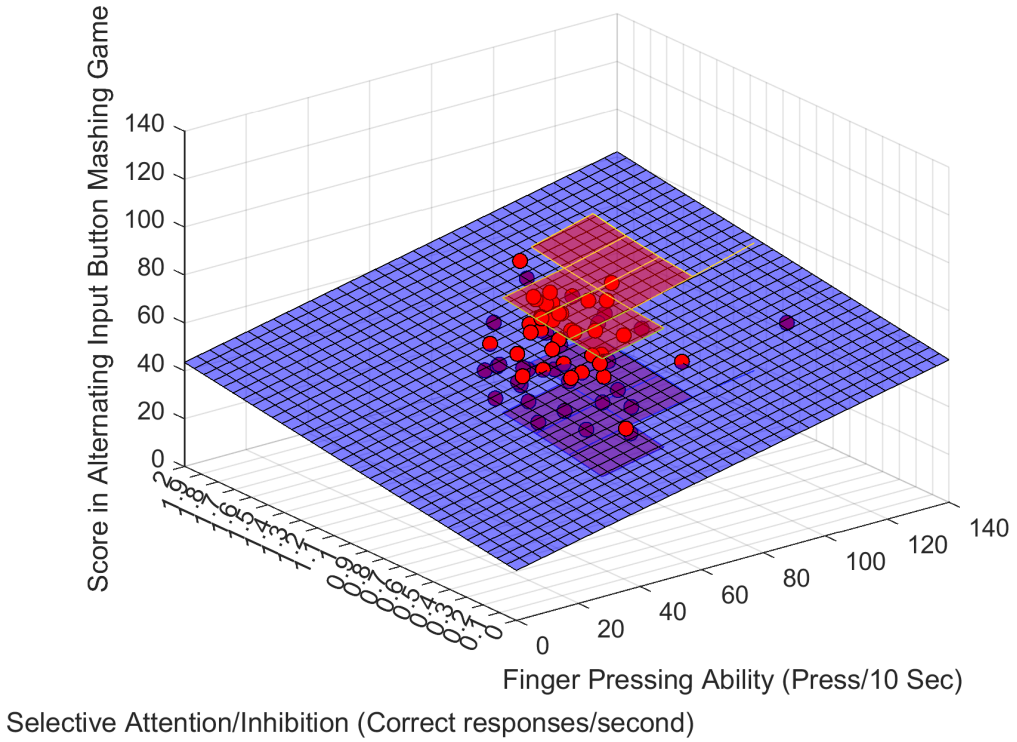
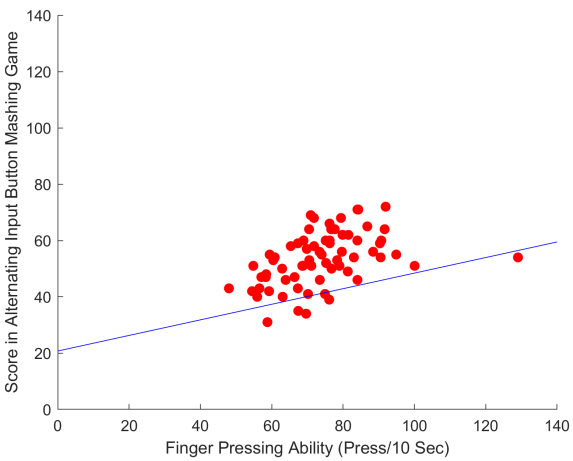
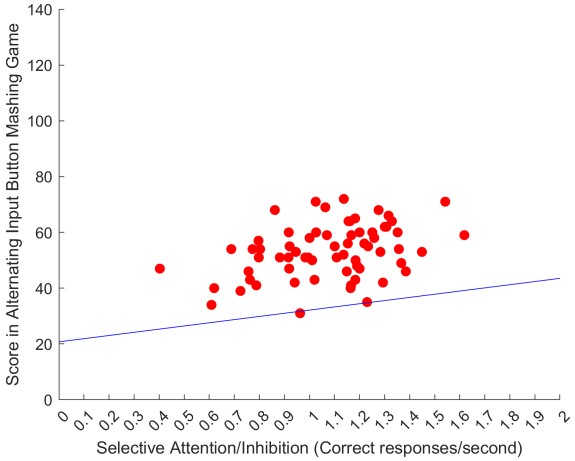


Figure 16.2: AIBM regression model plotted with the raw data.



(a) Data (red dots) on a slice of the AIBM model plane where SelAtt = 0 (blue line).



(b) Data (red dots) on a slice of the AIBM model plane where FP = 0 (blue line).

Figure 16.3: Slices of AIBM Model plane where the predictor variables are held at 0 to compare their relationship to the data.

Regression Modeling:: MIBM Score $= 28.38 + 0.55 \times$ **FP**

The MIBM regression model is significant: $F_{(1,66)} = 21.96$, $p < 0.0001$, $R^2 = 0.25$(details in Table 16.5). FP is a significant predictor ($t(66) = 4.69$, $p < 0.0001$) with each additional press per 10 seconds increasing the player's score by 0.55 units.

Variable	Coefficient	CI (95%)	Std. Error	t-stat
Finger Pressing	0.55****	[0.32, 0.79]	0.12	4.69
Intercept	28.38***	[10.65, 46.10]	8.88	3.20
	$R^2 = 0.25$	$F_{(1,66)} = 21.96$****		
Significance: not significant ( ), $p < 0.05$(*), $p < 0.01$(**), $p < 0.005$(***),$p < 0.001$(****)				

Table 16.5: Stepwise regression model for MIBM.

Plotting the regression line against the data we see a handful of data points that seem quite far below the model (Fig. 16.4). Based on their location, these data points seem to represent participants who are greatly underperforming for their ability level. This is likely to have a strong effect in the residuals that would warrant further investigation.

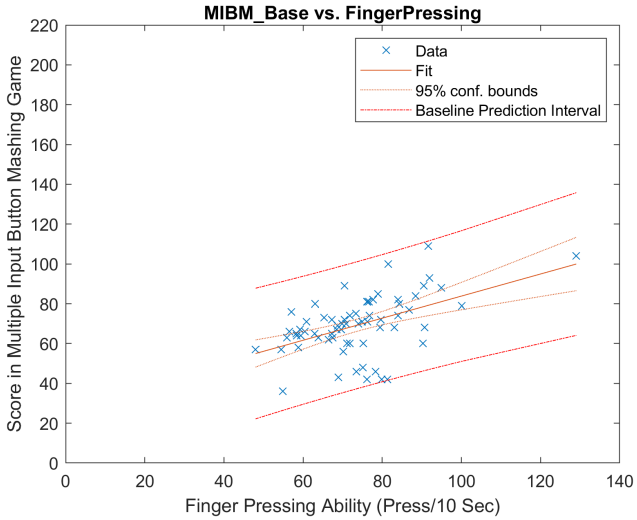


Figure 16.4: MIBM regression model plotted with the raw data.

16.5 Checking Model Assumptions

Table 16.6 summarizes our comparison of the models to assumptions from Sec. F.0.2.

	Assumptions			
Model	**Correlation**	**Collinear**	**Normality**	**Homoscedastic**
SIBM	✔	✔	✔	✔
AIBM	✔	✔	✔	✔
MIBM	✔	✔	✖	✖
✔ indicates meets assumptions, ✖ indicates not meeting assumptions				

Table 16.6: Summary of whether models meet the regression assumptions.

16.5.1 Correlation and Collinearity

Correlation: Table 16.7 summarizes correlations between variables. Our resulting models only contain IVs that are significantly correlated with DVs, and so they all meet this assumption.

	Correlation Coefficients for Performance in...		
Ability	**SIBM**	**AIBM**	**MIBM**
Finger Pressing	0.74***	0.44***	0.50***
Selective Attention	0.15	0.37**	0.1095
Inhibition	0.15	0.11	0.04
Object Recognition	0.10	0.30*	0.17
Token Change Detection	-0.10	0.04	0.002
Significance: not significant ( ), $p < 0.05(^*)$, $p < 0.005(^{**})$, $p < 0.0005(^{***})$			

Table 16.7: Correlational Coefficients (Pearson r) for Abilities and Performance Score in the three challenges.

Collinearity: Table 16.8 summarizes the correlations between the IVs and their VIFs (in the diagonal). SIBM and MIBM meet this assumption as they do not have multiple predictors. AIBM also meets this assumption as its two variables (FP and SelAtt) are not significantly correlated (0.22, p = n.s.), and their VIFs are low (≤ 3).

	FP	**SelAtt**	**Inhib**	**OR**	**TCD**
FP	(1.08)	0.22	0.19	0.14	-0.05
SelAtt		(1.82)	0.45***	0.55***	0.25*
Inhib			(1.29)	0.20	0.19
OR				(1.44)	0.15
TCD					(1.10)
Legend:	Finger Pressing (FP), Selective Attention (SelAtt), Inhibition (Inhib), Object Recognition (OR), Token Change Detection (TCD)				
Significance: not significant ( ), $p < 0.05(^*)$, $p < 0.005(^{**})$, $p < 0.0005(^{***})$					

Table 16.8: Pairwise correlation matrix for independent variables with VIF values in the diagonal.

16.5.2 Normality and Homoscedasticity

We evaluate these assumptions through interpreting plots. Since this is interpretive, we are conservative in saying the models meet assumptions. We find SIBM and AIBM meet the normality and homoscedasticity assumptions, but MIBM does not.

SIBM

Homoscedastic: We plot standardized residuals versus the fitted values (Fig. 16.5). We see the data is randomly distributed around the 0-line, with no obvious patterns. We notice at least three significant outliers which we keep in the dataset as they do not seem unreasonable in their values considering human variation. Overall this implies residual independence and homoscedasticity.

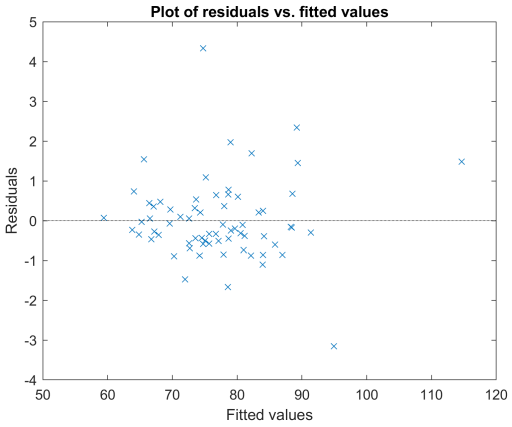
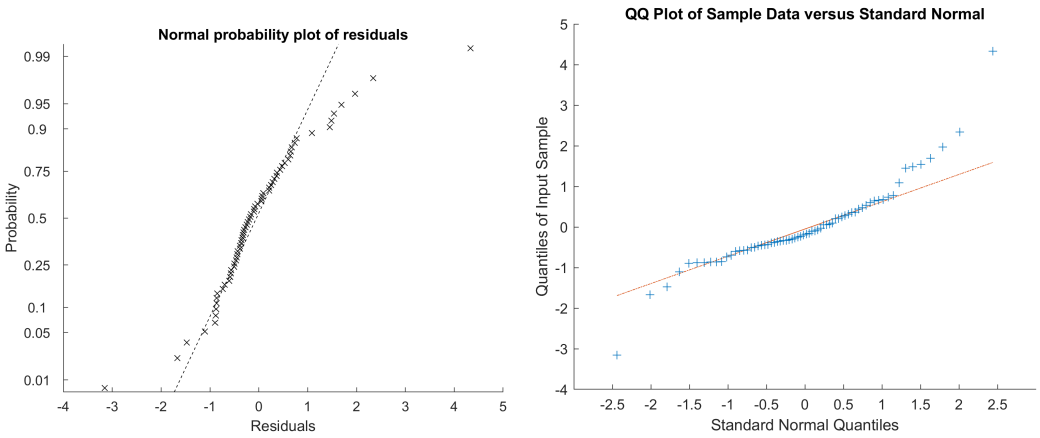


Figure 16.5: Standardized Residuals vs. Fitted Values Plot for SIBM.

Normality: The P-P Plot (Fig. 16.6a) shows the data looks normal around the centre of the distribution[12], with some slight bulging that could indicate the data is more peaked than normal. Checking the Q-Q Plot (Fig. 16.6b) we see that the data curves upwards indicating that it may be heavily-tailed. Both of these could be the result of the homogeneous participant sample (able-bodied, young, gamers). This demographic is likely to be high performers in ability measures and game scores, and so would exist at the peak of a proper normal distribution or towards the right-tail in the raw scores. With many data points falling close together, a well fit model would end up with peaky residuals, and likely underestimate any data points that fall significantly far above the mean causing tailing.



(a) P-P Plot (b) Q-Q Plot

Figure 16.6: Probability plots for SIBM Standardized Residuals with Normal reference.

[12]P-P Plots have more power around the centre and are fuzzy near the edges.

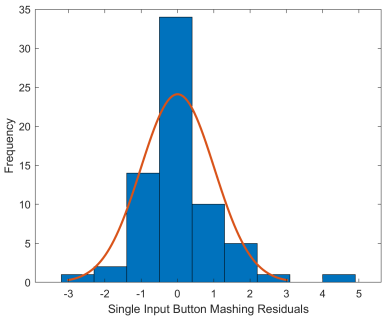


Figure 16.7: SIBM Histogram with normal distribution reference.

The histogram of standardized residuals (Fig. 16.7) confirms the strong peaking and tail on the right side. This implies we are on the right track with the demographic explanation. Given this sampling information, it seems like the normality does not deviate significantly from our expectations to make us question the validity of the regression.

AIBM

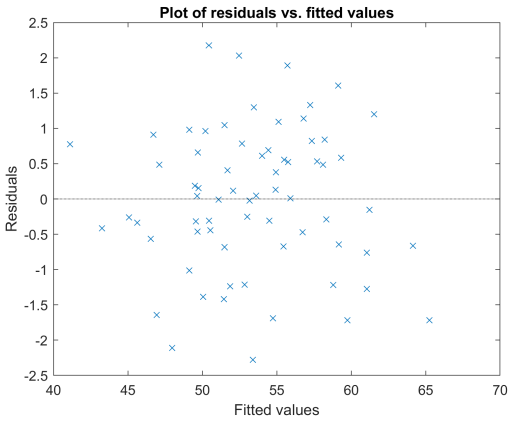
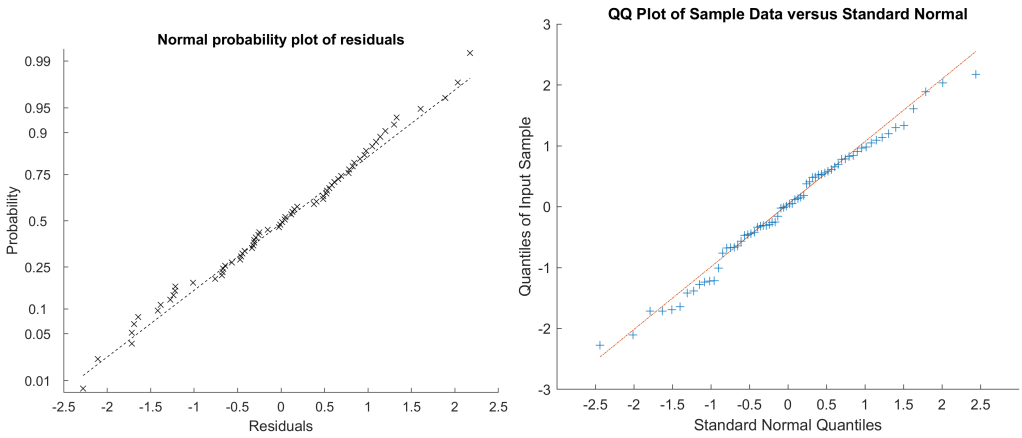


Figure 16.8: Standaradized Residuals vs. Fitted Values Plot for AIBM.

Homoscedastic: The standardized residuals versus fitted values plot (Fig. 16.8) shows a random scattering around the 0-line. No points stand out as obvious outliers, and the points are quite distributed (no obvious banding). The residuals do not seem to be autocorrelated (*Durbin-Watson = 1.98*). Overall this implies that the residuals are independent and homoscedastic.

Normality: The P-P Plot (Fig. 16.9a) is fairly consistent with the reference line around the centre. The Q-Q Plot (Fig. 16.9b) similarly shows the edges are fairly consistent with the reference line. Looking closer at both plots we see some interesting jumps/ gaps in the data that are almost unnoticeable at first glance. It is possible these gaps are due to noise in the data, or a latent variable. However, the fact these plots are so close to the normal reference line implies that even if this is the case it does not have a significant effect to make the data bimodal.



(a) P-P Plot (b) Q-Q Plot

Figure 16.9: Probability plots for AIBM Standardized Residuals with Normal reference.

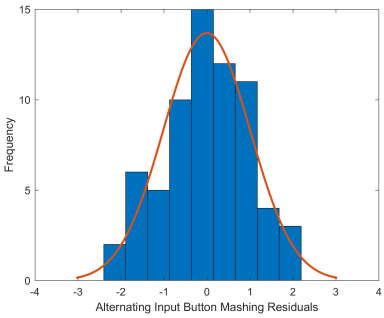


Figure 16.10: AIBM Histogram with normal distribution reference.

Confirming with the histogram (Fig. 16.10) shows a fairly normal distribution, with mild peaking around the centre and at some points in the tails. However, as a whole it is fairly balanced (no strong tails). As such we are not concerned about the data, and believe the residuals meet the assumption of normality well.

MIBM

Homoscedastic: The standardized residuals versus the fitted values (Fig. 16.11) seem to have patterns, though not singular or cohesive enough to indicate a non-linear model. There are obvious outliers significantly above and below the 0-line. We do see more points above the 0-line, which could indicate the model is underestimating performance. The outliers on the positive residual side seem to increase linearly, but their meaning is unclear as the rest of the positive side residuals are fairly random. Many points on the negative side group around -2 or lower, indicating model overestimation. Overall we cannot say the residuals meet independence and homoscedasticity.

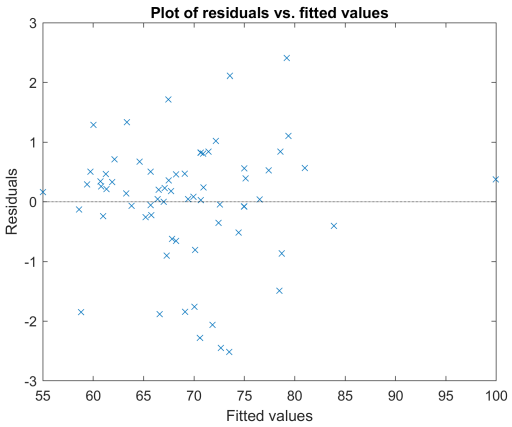
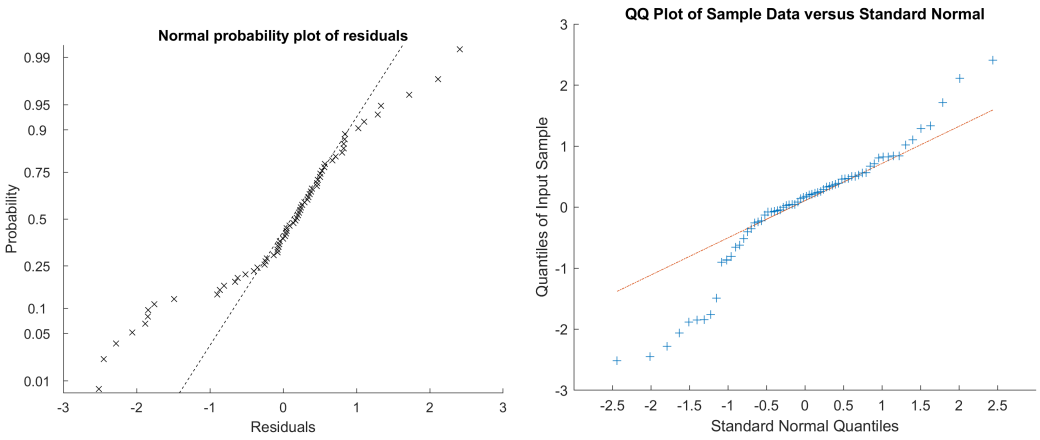


Figure 16.11: Standaradized Residuals vs. Fitted Values Plot for MIBM.

Normality: The normal probability plots (Fig. 16.12) show significant deviation from the normal reference around the tails. The P-P Plot (Fig. 16.12a) shows the data around the centre seems to fit a normal distribution (no significant deviations from reference). The line begins to significantly diverge on the left side (< 0) and in the tail on the right side (> 1). The Q-Q Plot (Fig. 16.12b) indicates the data may be heavy-tailed from the shape. The way the data is bunched in the positive residuals and quantiles implies the model is underestimating performance (since most points exist in positive residual). From the Q-Q Plot there seems to be a gap in density around Normal Quantile -1, and a significant number of values in the lower tail. In this case, it seems like the model is significantly overestimating these participant's performance.



(a) P-P Plot (b) Q-Q Plot

Figure 16.12: Probability plots for MIBM Standardized Residuals with Normal Distribution reference line.

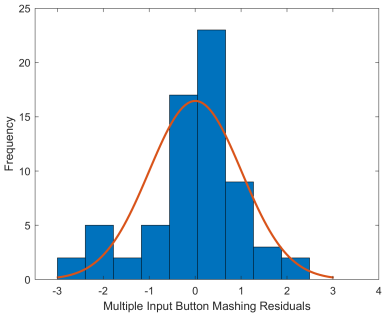


Figure 16.13: MIBM Histogram with normal distribution reference.

Looking at the histogram of standardized residuals (Fig. 16.13) we see some peaking between 0 and 1. We also notice the difference between the distribution and frequencies on the left side of the histogram, and strong left-tail. This seems consistent with what we are seeing in the normality plots. Considering the model against the data (Fig. 16.4) it seems like there was a group of significantly underperforming participants. This would explain the issues with the residuals and normality, as the model would be trying to fit all these points in the same regression and so would be significantly affected by these underperformers. It is unclear at this time *why* this occurred and so it does not make sense to remove them from the model.

16.6 Discussing Model Performance

The goal of our study is to validate our competency profiles. The regression models represent the "real" relationship between abilities and button mashing scores which we loosely compare our competency profiles against. Hence the estimation/predictive power of specific regression models is not sufficiently important at this time. Our models are currently quite uncertain, as indicated by their prediction intervals and R^2 values (AIBM, MIBM $< 30\%$, SIBM 54%)[13]. However, the **elements that affect our model's fits could represent latent variables in this relationship**. We discuss two potential impacts across all models: human variability and participant strategy. We also discuss two more impacts specific to MIBM: the combination of handedness and movement, and latent abilities.

[13]While the R^2 is low, the statistical power is high. Given the sample size (N=68), $p < 0.05$, and their respective R^2's, their statistical powers are: SIBM = 1.0 (1 predictor) AIBM = 0.9957 (2 predictors), MIBM = 0.9971 (1 predictor).

Human variability: Natural variation between tests, and fatigue could explain why we had sudden underperformers in MIBM, or the varied number of outliers in each game. Though each mini-game is only ten seconds, the repetitive button mashing is extremely fatiguing on hands and joints. While we randomized the order of presentation, fatigue affects every player differently and so the results may be more noticeable in some areas (i.e. through significant underperforming outliers).

Strategy: We notice participants use different action strategies during play (Table 16.9). One-hand participants use two fingers on the same hand, frequently moving at the metacarpophalangeal joint (first knuckle). Two-hand participants use a finger from each hand, and move either at the finger joint, wrist, elbow or shoulder. Looking at the distributions for AIBM and MIBM, this may explain deviations. We see a gap in AIBM's probability plots (Fig. 16.9) which could indicate bimodality in the data; could handedness be this separation? If this is the case, it seems to not affect the overall residual distributions enough to have large effects on the model. Looking at MIBM's data (Fig. 16.4), we see 16 "underperformers" (points below the confidence bounds) which coincidentally is the number of two-handed participants. However, our observational notes state that these underpeforming participants use a mixture of one and two handed approaches, making it unlikely this gap is exclusively to do with handedness. Our observational notes indicate these underperformers seem to use a variety of movements (e.g. moving at shoulder, vs finger).

*N = 67**	**SIBM**	**AIBM**	**MIBM**
One-handed	65 (96%)	39 (57%)	51 (75%)
Two-handed	2 (3%)	28 (41%)	16 (24%)

*Note: We were unable to note the handedness of one participant during their baseline measure due to managing multiple participants.

Table 16.9: Handedness distribution per button mashing game.

Handedness and Movements: For MIBM specifically, a combination of handedness and movement location could create circumstances where participants with "good" FP levels would receive "bad" scores because they are not adequately synchronizing their presses. The experimental software is forgiving in what counts as "simultaneous" so long as the player is not accidentally alternating. It is possible that some of the underperformers were releasing the first button before the second was fully depressed (creating an alternating motion). Participants using certain two-handed strategies may be prone to falling out of sync with themselves in this way. Since we measure FP as a single button input skill, it could be the case that techniques and movements that create "good" FP lead to bad habits in the simultaneous pressing games which creates a "bad" MIBM score.

Latent ability: MIBM's poor fit and probability plot gaps imply there is a missing variable from the model that cannot be reasonably explained with strategy. This makes sense considering our close readings identified a cognitive ability would be necessary to synchronize

movements. It is possible our cognitive measures for SelAtt and Inhib (which we hypothesized as part of MIBM) are not detailed or granular enough to show up as part of this model. This is a reasonable concern given the mini-games used to measure these abilities were not clearly validated in Ch. 11. It is also possible that an unknown ability we did not measure is at play here.

16.7 Comparing Models to Competency Profiles

Table 16.10 summarizes our comparison of competency profiles to models via H1 and H2.

		Competency Profiles	
Game	**Model**	**H1 (Abilities)**	**H2 (Importance)**
SIBM	$26.70 + 0.68 \times \text{FP}$	✔	✔
AIBM	$20.75 + 0.28 \times \text{FP} + 11.38 \times \text{SelAtt}$	✔	✔
MIBM	$28.38 + 0.55 \times \text{FP}$	✖	✖
Legend: Finger Pressing (FP), Selective Attention(SelAtt)			

Table 16.10: Comparing Regression Models to Competency Profiles Summary

We can visually assess the hypotheses by plotting the model's standardized coefficients as a competency profile and comparing the two plots. We calculate each model's standardized coefficients from their unstandardized coefficient, using the equation:

$$\beta_X = \text{B}_X \times \frac{\sigma_X}{\sigma_Y}$$

Where X is an ability, Y is the game score, B_X is the X's unstandardized coefficient, σ_X is the standard deviation of X's raw values, and σ_Y is the standard deviation of Y's raw values.

A Quick Aside...

Recall the competency profiles combine SelAtt and Inhib since they are highly coupled abilities. The Mini-game Ability Battery's Looking and Recipe game both measure combined SelAtt and Inhib. For the purposes of modeling, each game was assigned one of these abilities (Looking: SelAtt, Recipe: Inhib). Therefore when either of these abilities show up in the regression models, they should be thought of as both abilities. The regression model will be considered to have met the SelAtt/Inhib abilities if either of those are in the model.

16.7.1 SIBM

From Fig. 16.14 we see that the only ability that needed to be in the regression model was FP (Rank 4). Since it includes FP and by default preserves the order of importance the SIBM competency profile and model meet H1 and H2.

$$26.70 + 0.68 \times \text{FP}$$

Ability	B	β
FP	0.68	0.74

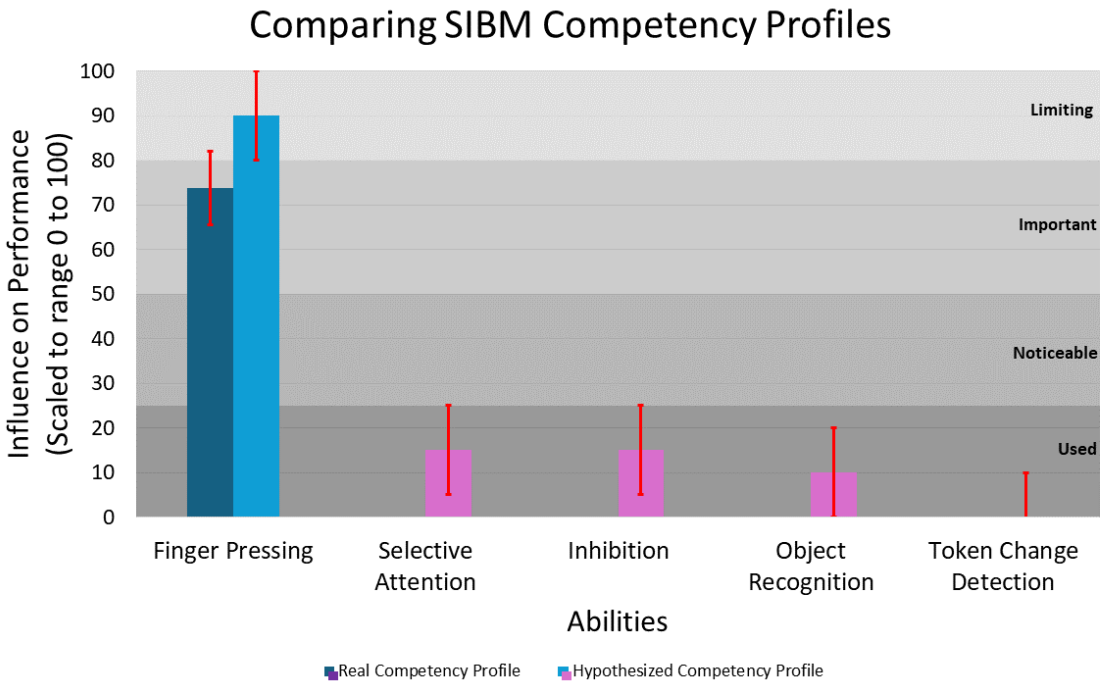


Figure 16.14: Comparing SIBM Regression Model to Hypothesized Competency Profile.

16.7.2 AIBM

From Fig. 16.15 we see FP and SelAtt are in the model, so H1 is met. From the shape (i.e. standard coefficients) the order seems preserved. We double check this against the regression steps: FP enters the model first ($F = 16.19$, $p < 0.001$), then SelAtt ($F = 6.69$, $p = 0.011955$). This aligns with what we expect from the competency profile, and so H2 is met.

16.7.3 MIBM

MIBM does not meet either hypothesis. This may be because the MIBM does not meet our reasonable regression assumptions. From analysing the model, we suspect MIBM to have a latent variable. Considering our competency profile proposes SelAtt/Inhib as being borderline between Rank 1 and 2, it is possible this is the variable and our existing measures were not sufficient to pick up on it.

16.8 Conclusion from the Study

Our results show that our competency profiles are reasonable approximations for ability requirements of button mashing challenges. The single input button mashing (SIBM) and

$$20.75 + 0.28 \times \text{FP} + 11.38 \times \text{SelAtt}$$

Ability	**B**	β
FP	0.28	0.38
SelAtt	11.38	0.28

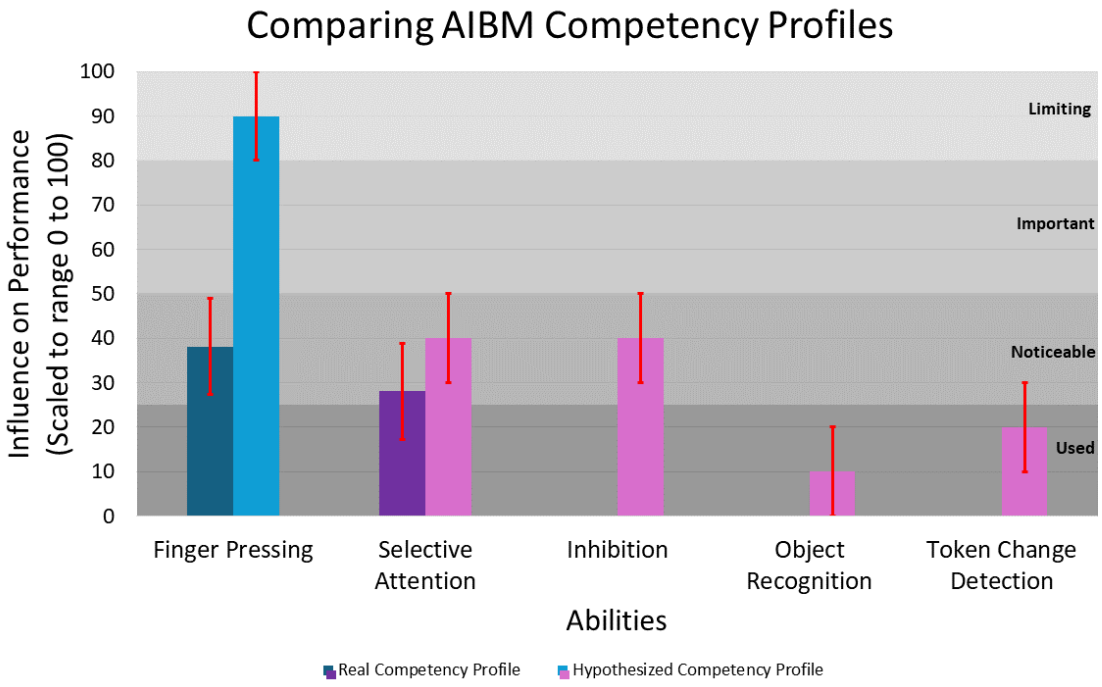


Figure 16.15: Comparing AIBM Regression Model to Hypothesized Competency Profile

alternating input button mashing (AIBM) models aligned well with our competency profiles and met our hypotheses. AIBM surprisingly shows a significantly smaller finger pressing importance than we anticipated, though this could be due to our overestimation in the competency profile or our small sample size in this study. The multiple input button mashing (MIBM) model was not reasonable, and did not match our competency profile. However, its diagnostics seemed to indicate that it was missing a variable which aligns with what we hypothesized in our competency profile. Overall these results suggest our competency profile construction process is sufficient, and that with a larger more focused study we could pinpoint the issues with MIBM.

This is an exploratory study subject to limitations outlined in Sec. 15.5. To get a more robust idea of which abilities are used in button mashing games a larger scale experiment with a factor analysis method would be more appropriate. Moving forward from this point we need to consider how to compare these regression models to player profiles.

$$28.38 + 0.55 \times \text{FP}$$

Ability	B	β
FP	0.55	0.50

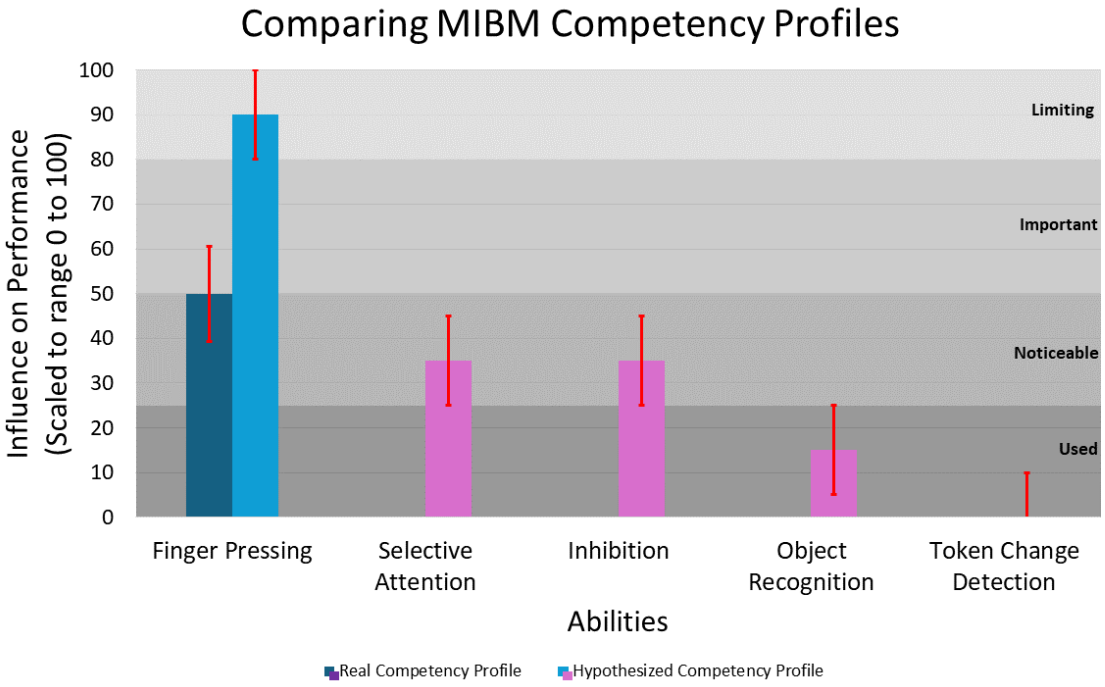


Figure 16.16: Comparing MIBM Regression Model to Hypothesized Competency Profile.

Take home points

From this chapter we learned the following meta-lessons:

- Competency profile construction method seems reasonable given 2 competency profiles (SIBM and AIBM) seem to be accurate; 1 competency profile (MIBM) seems reasonable, but could not be confirmed.
- Limitations in Ability Battery measurements could be hiding relationships with dependent variables.
- Participant strategies and gaming experience may play a larger role in performance at atomic challenges than anticipated.
- Regressions (and data) suffer from not being more strict in the participant sample and their approaches to the games.

Chapter 17

(Over)Loading and Competency Profiles (Study 1B)

We conduct an exploratory study to understand how under and over-loading limiting abilities via mechanical difficulty affects a challenge's competency profile. In this process we provide more evidence for our baseline models (Ch. 16), and address RQ2.

We begin this chapter explaining the relationship between load and performance. We lay out our hypotheses and connect them to the button mashing challenges. We summarize our study design, participant information (for the set and each group), and data analysis process. We then present our results, which are the majority of this chapter's content. We conclude by discussing how the competency profiles change given the adjusted mechanical difficulty, and how future work could further explore this topic.

17.1 Loading and Performance

"Mental workload" describes the demand a task places on a person's cognitive resources [242][1], though it can encompass motor responses as part of the information-processing model. Low-demand tasks (i.e. *underload*) are easy to complete, but high-demand tasks (i.e. *overload*) cause performance break down [517]. Wicken's Multiple Resource Theory models overall workload as the sum of individual ability loadings [517]. ***Challenge competency profiles are effectively workload representations***, as they quantify each ability's loading and identify the limiting ability[2].

To explore the relationship between competency profiles (*workload as individual ability loadings*) and challenge achievability (RQ2) we need to manipulate task demand. Bowman and Tamborini [54] adjust task demand via changing the number of actions required to play the game. We take a similar approach by manipulating *mechanical difficulty* to change the load of the limiting ability. Mechanical difficulty allows us to map challenge variables to abilities, such that we can manipulate a game mechanic to affect specific abilities in the competency profile. For our button mashing games, we manipulate load for the limiting

[1]Kosch et al. [242] note that this is a common understanding of workload in HCI, though it is often mixed with the idea of "cognitive load" from cognitive load theory[460].

[2]Ability which contributes most to the player's performance in the game.

ability via score modifiers (see Custom Games in Ch. 15.2). The score modifier is a source of mechanical difficulty that changes what "counts" as a correct input. As a proxy for Finger Pressing (FP), the relationship between the score modifier and FP load is inverted: a higher score modifier underloads FP (fewer presses needed), while a lower score modifier overloads FP (more presses needed).

17.1.1 Hypotheses

Hypotheses:

Since we are scaling limiting ability load via score modifier, we expect...

H1: Performance in the games will scale constantly based on the score modifier.

H2: Competency profiles that rely only on the limiting ability will scale constantly from their baseline model by the score modifier.

H3: Competency profiles that rely on more than the limiting ability will see their limiting ability scale by the modifier, but their secondary abilities will not scale neatly.

Since the score modifier constantly changes what "counts", we expect that performance in all games will be constantly scaled up or down based on the condition (H1). We expect Single Input Button Mashing (SIBM) and Multiple Input Button Mashing (MIBM) will fit H2, such that underloading produces scores twice as large for half as much FP as the baseline and overloading produces scores half as large but requires almost twice as much FP. However, we acknowledge that the MIBM competency profile from the baseline study was inconclusive and so the changes may be different than we expect. We think Alternating Input Button Mashing (AIBM) fits H3: performance will be scaled with the score modifier, so the limiting ability load will change accordingly. However, we think selective attention (SelAtt) will become more important when FP is overloaded, and less important when FP is underloaded, due to its supporting role in the task. We expect this to show up in the competency profile as dramatic differences in SelAtt importance between conditions.

17.2 Study Design

We conduct a single-factor, multi-level experiment with between-groups design to view each difficulty condition separately and compare them against each other. We have three conditions: Easy, Hard, and Control (i.e. baseline model tuning). Details for the study design, procedures, apparatus are in Ch. 15.

17.2.1 Participants

We remove two participants (P05 and P62) due to technical errors in capturing their FP data (N= 73; 46 men, 23 women, 4 queer[3]). We assign participants to experimental conditions via block randomization, leading to three groups similar in size, age, gender distribution, and ability scores (see Table 17.1).

[3]Queer is used as shorthand for participants who identify as non-binary, trans, queer, or unidentified.

	N	$\overline{\text{Age}}$	M:W:Q	$\overline{\text{FP}}$	$\overline{\text{SelAtt}}$
Easy	23	22.0	14:8:1	76.52 (σ 17.63)	1.02 (σ 0.20)
Control	25	21.6	14:8:3	74.81 (σ 11.22)	1.13 (σ 0.22)
Hard	25	22.1	18:7:0	73.85 (σ 12.46)	1.08 (σ 0.28)
***Legend*:**	$\overline{X}$: the mean of X, N: number of participants, M: Man, W: Woman, Q: Queer, FP: Finger Pressing, SelAtt: Selective Attention				

Table 17.1: Descriptive information for each condition.

17.2.2 Analysis Method

All statistical processes and visualization are done via Matlab 2024a [283] using the Statistics and Machine Learning Toolbox [282]. We summarize our analysis processes in Fig. 17.1. We have two goals in this study: validate the baseline models from Ch. 16, and explore the competency profiles for different loadings. We approach both goals by first checking data distributions, and then running appropriate regressions.

Data Distributions: We use box-and-whisker plots to compare predicted and actual score distributions. This explores whether player performances were roughly "the same" between the baseline and experiment. For each condition, we expect the participant's actual score is roughly their baseline score multiplied by the condition's score modifier. Small differences from noise like human variation may occur. We decide a mean difference within ± 10 points is small enough to be considered the effect of human variation.

Regressions: We run regressions for two purposes: finding models for different loadings, and comparing loading models. We compare conditions, and predicted vs. actual models, by running regressions with the condition (either Easy/Control/Hard or Predicted/Actual) as a categorical variable.

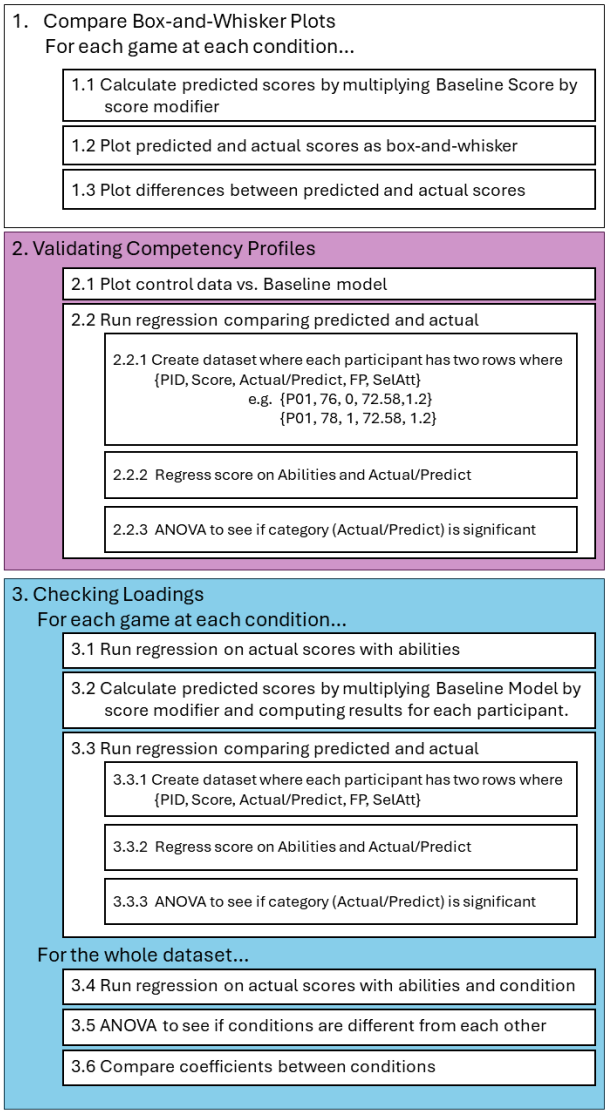


Figure 17.1: Data Analysis Procedure for Loading Study.

The models are "different" if the interaction terms between the predictors and categorical variables are significant. We have a small number of observations per condition (Table 17.1). For SIBM and MIBM regressions which only use a single predictor, these numbers may be sufficient (though it is possible we may see deviations from normality caused by this smaller sample size). However, AIBM's low numbers are *just* sufficient for regression with two predictors. While our work is exploratory and so will proceed under this view, we want to make it clear that these limitations may come into play. We report the regression assumptions and diagnostics for the results in App. G.1.

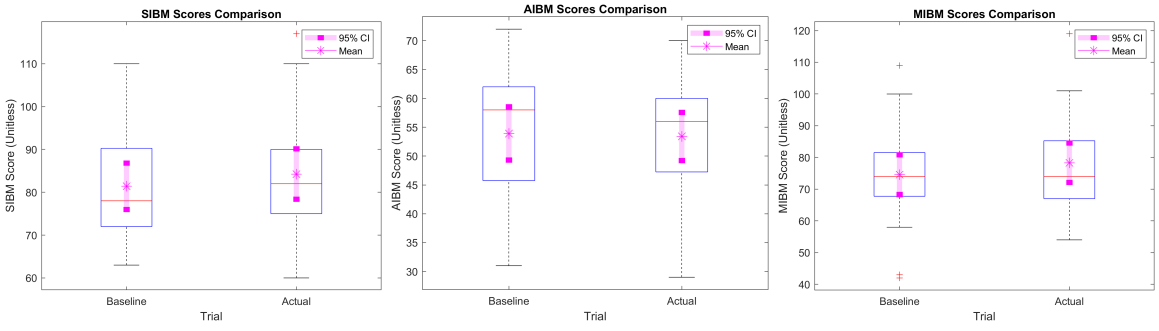
A Quick Aside...

The analysis for this study uses both statistical tests and visual inspections/interpretations of plots. The language surrounding visual interpretations may be less precise than the language used to describe results of statistical tests. We make sure to clarify in the body when we are drawing a result from visual data inspection, so as to prime readers for our interpretive language. Where possible, we present statistical data alongside the interpretive description to give the reader more context. We also make sure to reference the figures being discussed, such that readers can make their own judgment on whether they agree with our interpretation.

17.3 Results: Validating Baseline Models

We use the Control condition (N= 25) to test our baseline regression models from Ch. 16. Validated baseline models allow for more meaningful comparison of the Easy and Hard conditions to the Baseline. **Overall we find the baseline models reasonably fit the control data for SIBM and AIBM. MIBMs distributions imply a reasonable fit, but checking via regression suggests otherwise.** We decide to continue with using MIBM for comparisons, even though they are unlikely to produce meaningful results.

17.3.1 Comparing Raw Experimental Scores to Raw Baseline Scores



(a) SIBM (b) AIBM (c) MIBM

Figure 17.2: Comparison of Predicted Scores to Actual Scores for Control Condition.

The box-and-whisker plots for SIBM (Fig. 17.2a), AIBM (Fig. 17.2b), and MIBM (Fig. 17.2c) imply the control condition's performance was almost identical between their experimental and baseline measures. There is some slight shifting into lower scores in the actual measure. The median lines are approximately the same, and the means (and their confidence intervals) overlap.

On visual inspection, the distribution of the differences between actual and predicted values for each game further suggests the baseline and control sets are effectively the "same" (Fig. 17.3). The confidence intervals for each conditions' mean difference all include 0 in their range (SIBM: [-0.40, 6.08], MIBM: [-0.23, 7.75], AIBM: [-1.92, 0.88]), implying no significant difference between the groups. The experimental scores for AIBM and MIBM are within 10-points from their respective baseline scores (indicated by the dotted reference lines). We see some larger variability in SIBM's differences in the positive direction, which indicates the actual score is higher than the predicted score for some participants.

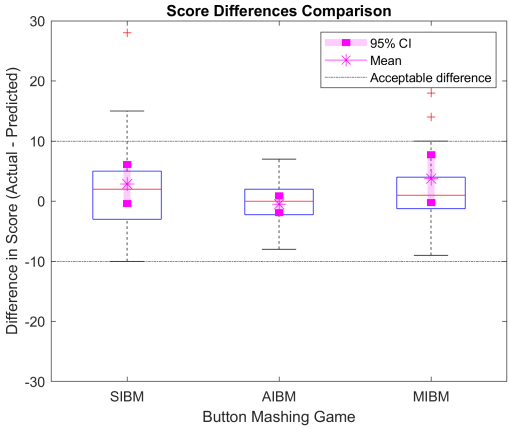
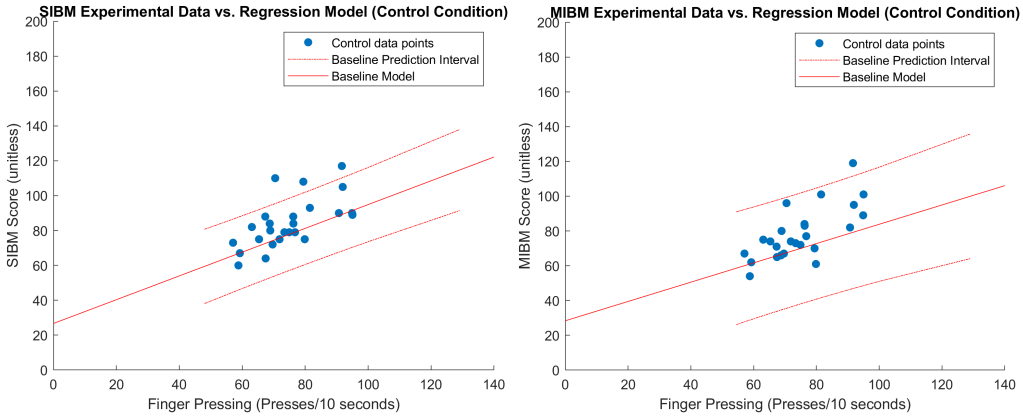


Figure 17.3: Score Differences for Control Condition.

17.3.2 Checking Control Scores Against Baseline Model



(a) SIBM (b) MIBM

Figure 17.4: Control Data plotted against Baseline Regression Models

We plot the 95% simultaneous observation prediction intervals for the baseline model to see whether the control data fits into the range we would expect. For both SIBM (Fig. 17.4a) and MIBM (Fig. 17.4b), many data points lie above the regression line, but fall within the prediction intervals implying the model reasonably fits the new data. AIBM data points seem to be dispersed both above and below the regression plane (Fig. 17.5). Similar to the other games, all new observations fall between the predicted intervals.

We generate Baseline predictions for the group. We regress the Baseline and Control data on the appropriate abilities (SIBM: FP, MIBM: FP, AIBM: FP, SelAtt) with the data source

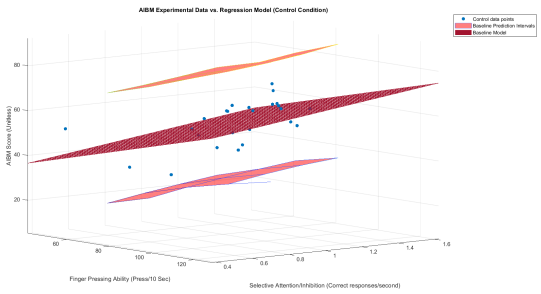
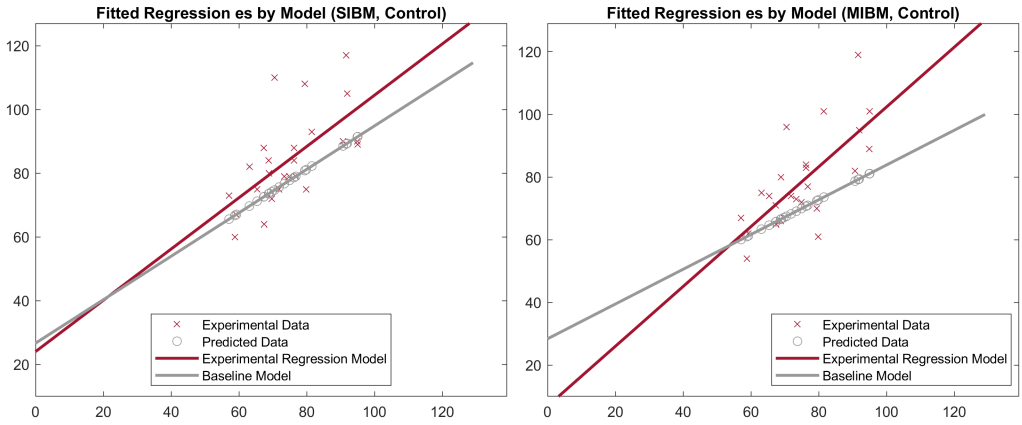


Figure 17.5: AIBM Control Data plotted against Baseline Regression Model.

(Experimental or Baseline-generated) as a categorical variable. Visually, SIBM's regression lines are close (Fig. 17.6a), and an ANOVA shows no statistically significant interaction between FP and the data source ($F_{(1,46)}$=*0.36, p=0.55*). MIBM's lines look significantly different (Fig. 17.6b). An ANOVA shows significant interaction between FP and data source ($F_{(1,46)}$=*4.13, p=0.048*), though just barely. This suggests that MIBM's control regression and Baseline model are incompatible, even though the control data fits within the prediction intervals.



(a) SIBM (b) MIBM

Figure 17.6: Comparing Baseline and Control Regressions.

The AIBM planes (Fig. 17.7) look reasonably similar. An ANOVA shows no significant interaction between the data source and either FP ($F_{(1,44)}$=*1.71, p=0.20*) or SelAtt ($F_{(1,44)}$=*0.65, p=0.43*), suggesting our baseline model is not significantly different from the control regression model.

17.4 Results: Regression Models for Loads

We produce regression models for the Easy (N= 23) and Hard (N= 25) conditions. **All models are significant, but MIBM's Overload (Hard) model, and AIBM's Underload (Easy) model do not reasonably meet regression assumptions.**

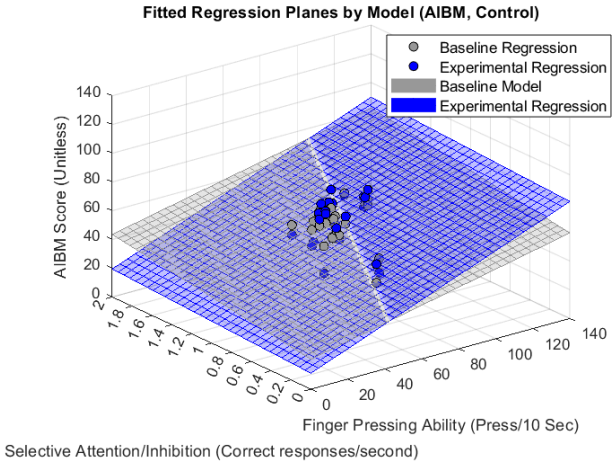


Figure 17.7: Comparing Baseline and Control Regressions - AIBM.

17.4.1 SIBM

Model	59.29 + 1.15 × Finger Pressing			
Variable	**Coefficient**	**CI (95%)**	**Std. Error**	**t-stat**
Finger Pressing	1.15****	[0.70, 1.61]	0.22	5.25
Intercept	59.29***	[23.46, 95.11]	17.23	3.44
	$R^2 = 0.57$	$Adj.R^2 = 0.55$	$F_{(1,21)} = 27.6$****	
Significance: not significant ( ), $p < 0.05$(*), $p < 0.01$(**), $p < 0.005$(***),$p < 0.001$(****)				

Table 17.2: Regression model for SIBM Easy Condition.

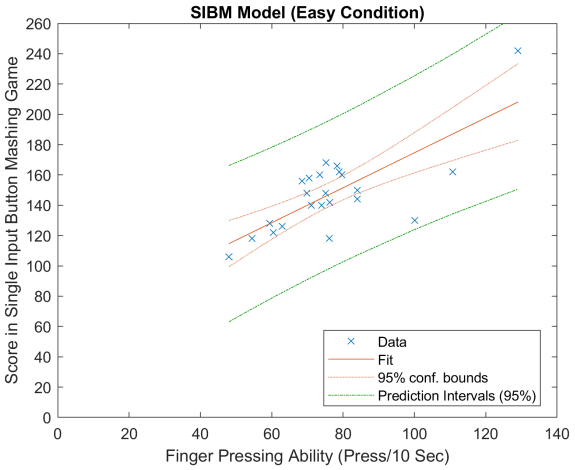


Figure 17.8: Plotted SIBM Regression Model for Easy Condition.

Easy (N=23). The regression results in a significant model, $F_{(1,21)} = 27.6$, $p < 0.0001$, $R^2 = 0.57$ (details in Table 17.2). FP is a significant predictor ($t(22) = 5.25$, $p < 0.0001$). Plotting the regression we see a positive linear trend (Fig. 17.8). There is a cluster of points above the confidence intervals of the regression line, however they fall within the prediction interval. This model seems to meet all regression assumptions (G.1.1) indicating it is a reasonable model for this game and condition.

[4] p=0.014503

Model	11.87 + 0.37 × Finger Pressing			
Variable	**Coefficient**	**CI (95%)**	**Std. Error**	**t-stat**
Finger Pressing	0.37****	[0.24, 0.49]	0.06	6.13
Intercept	11.87*[4]	[2.58, 21.16]	4.49	2.64
	$R^2 = 0.62$	$F_{(1,23)} = 37.5$****		
Significance: not significant ( ), $p < 0.05$(*), $p < 0.01$(**), $p < 0.005$(***),$p < 0.001$(****)				

Table 17.3: Regression model for SIBM Hard Condition.

Hard (N=25). The regression results in a significant model, $F_{(1,23)} = 37.5$, $p < 0.0001$, $R^2 = 0.62$ (details in Table 17.3). FP is a significant predictor ($t(23) = 6.13$, $p < 0.0001$). Plotting the regression we see a reasonably linear trend (Fig. 17.9). While most points fall inside the regression's confidence intervals, those outside seem to fall below the regression line more frequently than above. This model meets the correlation and auto-correlation assumptions, but tenuously meets the assumptions of residual normality, and homoscedasticity (see G.1.2 for details). For the purposes of this study we consider it as a reasonable model for this game and condition.

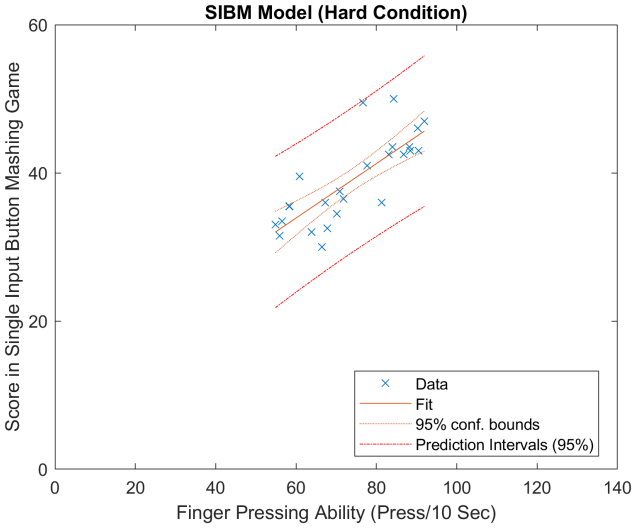


Figure 17.9: Plotted SIBM Regression Model for Hard Condition.

17.4.2 MIBM

Model	62.87 + 0.94 × Finger Pressing			
Variable	**Coefficient**	**CI (95%)**	**Std. Error**	**t-stat**
Finger Pressing	0.94***	[0.35, 1.53]	0.29	3.30
Intercept	62.87*[5]	[16.33, 109.41]	22.38	2.81
	$R^2 = 0.34$	$Adj.R^2 = 0.31$	$F_{(1,21)} = 10.9$***	
Significance: not significant ( ), $p < 0.05$(*), $p < 0.01$(**), $p < 0.005$(***),$p < 0.001$(****)				

Table 17.4: Regression model for MIBM Easy Condition.

[5] p=0.010514

Easy (N=23). The regression results in a significant model, $F_{(1,21)} = 10.9$, $p < 0.005$[6], $R^2 = 0.34$ (details in Table 17.4). FP is a significant predictor ($t(22) = 3.30$, $p < 0.005$). Plotting the regression (Fig. 17.10) we see a spread of data points on the right which fall below the regression line (although they are within the prediction interval). Investigating the regression assumptions for this model there seem to be some minor violations that may be explained by the small sample size (G.1.1). To this end we tentatively consider this model reasonable.

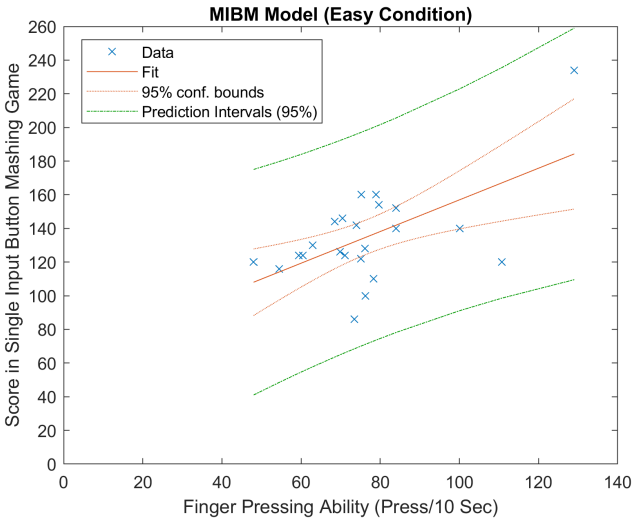


Figure 17.10: MIBM Regression Model for Easy Condition.

Model	16.03 + 0.27 × Finger Pressing			
Variable	**Coefficient**	**CI (95%)**	**Std. Error**	**t-stat**
Finger Pressing	0.27****	[0.13, 0.41]	0.07	4.01
Intercept	16.03***	[5.67, 26.38]	5.00	3.20
	$R^2 = 0.41$	$F_{(1,23)} = 16$****		
Significance: not significant ( ), $p < 0.05$(*), $p < 0.01$(**), $p < 0.005$(***),$p < 0.001$(****)				

Table 17.5: Regression model for MIBM Hard Condition.

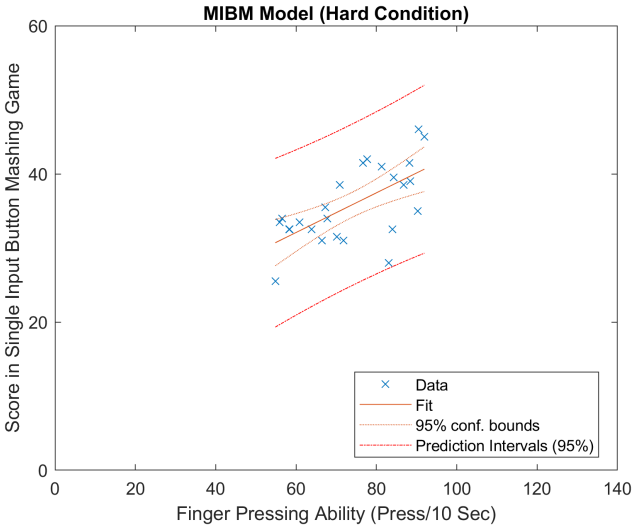


Figure 17.11: Plotted MIBM Regression Model for Hard Condition.

Hard (N=25). The regression results in a significant model, $F_{(1,23)} = 16$, $p < 0.001$, $R^2 = 0.41$ (details in Table 17.5). FP is a significant predictor ($t(23) = 4.01$, $p < 0.001$). Plotting the regression we see more curve-like trends than a linear one (Fig. 17.11). There are an almost equal number of points above and below the regression's confidence intervals. The model does not reasonably meet the assumption of residual normality and homoscedasticity (G.1.2).

[6] $p = 0.00343$

17.4.3 AIBM

Model	36.09 + 0.68 × Finger Pressing + 16.45 × SelAtt			
Variable	**Coefficient**	**CI (95%)**	**Std. Error**	**t-stat**
Finger Pressing	0.68**	[0.21, 1.15]	0.23	3.00
Selective Attention	16.45	[-24.77, 57.67]	19.76	0.83
Intercept	36.10	[-23.38, 95.56]	28.51	1.27
	$R^2 = 0.32$	$Adj.R^2 = 0.25$	$F_{(2,20)} = 4.59$*	
Significance: not significant ( ), $p < 0.05$(*), $p < 0.01$(**), $p < 0.005$(***),$p < 0.001$(****)				

Table 17.6: Regression model for AIBM Easy Condition.

Easy (N=23). The regression results in a significant model, $F_{(2,20)} = 4.59$, $p < 0.05$[7], $R^2 = 0.32$ (details in Table 17.6). FP is a significant predictor ($t(22) = 3.00$, $p < 0.01$), however SelAtt is not ($t(22) = 0.83$, $p = 0.414$). The regression plane increases steeply along the FP axis, with a shallower slope in the SelAtt axis (Fig. 17.12). The data falls between the prediction intervals, with many lying directly on or close to the main plane. This model is unreasonable as SelAtt and Scores do not significantly correlate, the residuals are heavily tailed, and the residuals-vs-fitted plot seems to have some clear patterning (G.1.1).

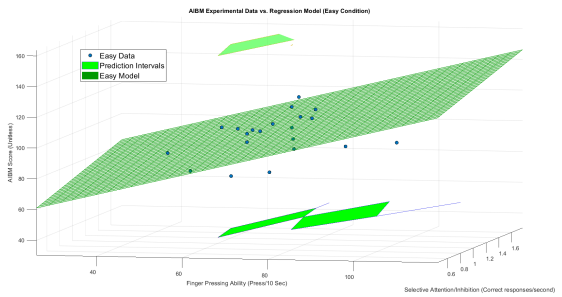


Figure 17.12: AIBM Regression Plots for Easy Condition

AIBM	9.28 + 0.22 × Finger Pressing + 2.23 × SelAtt			
Variable	**Coefficient**	**CI (95%)**	**Std. Error**	**t-stat**
Finger Pressing	0.22*[8]	[0.03, 0.41]	0.09	2.36
Selective Attention	2.23	[-6.36, 10.82]	4.14	0.54
Intercept	9.28	[-3.30, 21.86]	6.06	1.53
	$R^2 = 0.30$	$F_{(2,22)} = 4.78$*[9]		
Significance: not significant ( ), $p < 0.05$(*), $p < 0.01$(**), $p < 0.005$(***),$p < 0.001$(****)				

Table 17.7: Regression model for AIBM Hard Condition.

[7] $p = 0.0229$
[8] p=0.0279
[9] p=0.019

Hard (N=25). The regression results in a significant model, $F_{(2,22)} = 4.78$, $p < 0.05$[10], $R^2 = 0.30$ (details in Table 17.7). FP is a significant predictor ($t(22) = 2.36$, $p < 0.05$[11]), but SelAtt is not ($t(22) = 0.54$, $p = 0.595$). We see most of the datapoints are below the regression plane (Fig. 17.13). The model's residuals show minor violations in normality and homoscedasticity (G.1.2). However, we will cautiously consider this a reasonable exploratory model that could be improved with a larger sample size.

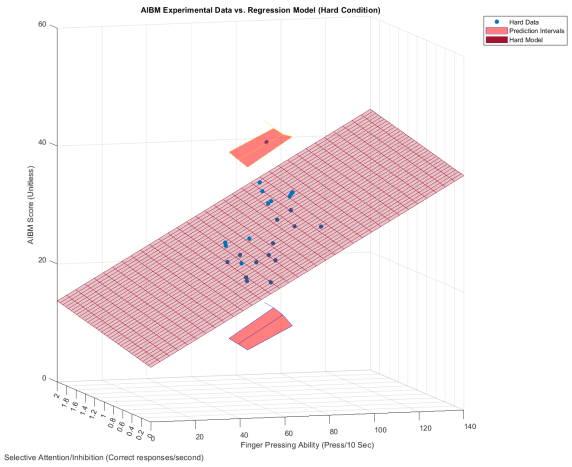


Figure 17.13: AIBM Regression Plot for Hard Condition.

17.5 Discussing Challenge Models

17.5.1 SIBM

Model Performance

Easy (Fig. 17.8). On visual inspection, the model seems to reasonably estimate player performance (supported by $R^2 = 0.57$). Minor over and under estimation seems reasonable given human variability, and the influence of outliers on the model. Three notable outliers fall below the regression line — these participants have high FP measures but very low scores. We do not believe that strategy plays a role in our under performing participants as there is only one person who used 2-hands in this condition. It is possible that participants in this condition were not "trying their hardest" and so we see unexpected variation. Alternatively these players could be succumbing to the effects of fatigue in their play.

Hard (Fig. 17.9). Visually, the model seems to estimate the player performance relatively well (supported by $R^2 = 0.61$), with most data points falling on or under the line. It is possible that participants scoring below the regression line are reaching their performance limit, either through fatigue or bio-mechanical speed limits. We also see interesting jumps in the data, where it appears there are multiple groupings of linear looking points. These jumps could indicate there is an latent variable. We do not think that strategy is the latent variable as all of the participants use a one-hand approach. However, there is the potential that the "strategy" is not how many hands, but rather the part of the body being moved (e.g. wrist, elbow, or shoulder). It could be that the latent variable is related to game literacy or experience in this way as well. The jumps that we see here are likely why the residuals for this regression were not perfectly normal, however it is hard to know whether this would go away with a larger sample size.

Control	24.06 + 0.80 × Finger Pressing							
Easy	59.29 + 1.15 × Finger Pressing							
Hard	11.87 + 0.37 × Finger Pressing							
Variable	**Coeff.**	**t-stat**	**p-value**	**SS**	**DF**	**MS**	**F-Stat**	**p-value**
Finger Pressing	0.80	3.59	0.0006	10058	1	10058	66.46	< 0.0001
Condition	–	–	–	135110	2	67554	446.38	< 0.0001
(Easy)	35.22	1.71	0.09	–	–	–	–	–
(Hard)	-12.19	-0.54	0.59	–	–	–	–	–
FP x Condition	–	–	–	1503.9	2	751.96	4.97	0.01
(Easy)	0.35	1.30	0.20	–	–	–	–	–
(Hard)	-0.44	-1.45	0.15	–	–	–	–	–
Model: $R^2 = 0.94$, Adj. $R^2 = 0.93$, $F_{(5,67)} = 203$, $p < 0.00001$								

Table 17.8: Regression and ANOVA results showing effects of model type for SIBM.

Comparing Regression Models

Are the conditions different? We regress the entire experimental dataset on FP with the condition as a categorical value. We use the Control condition as our reference point for interaction terms because we verify it is not statistically different from our baseline model. This results in a significant model ($F_{(2,67)}$=*203, $p < 0.00001$, R^2=0.94*), with the regression equation:

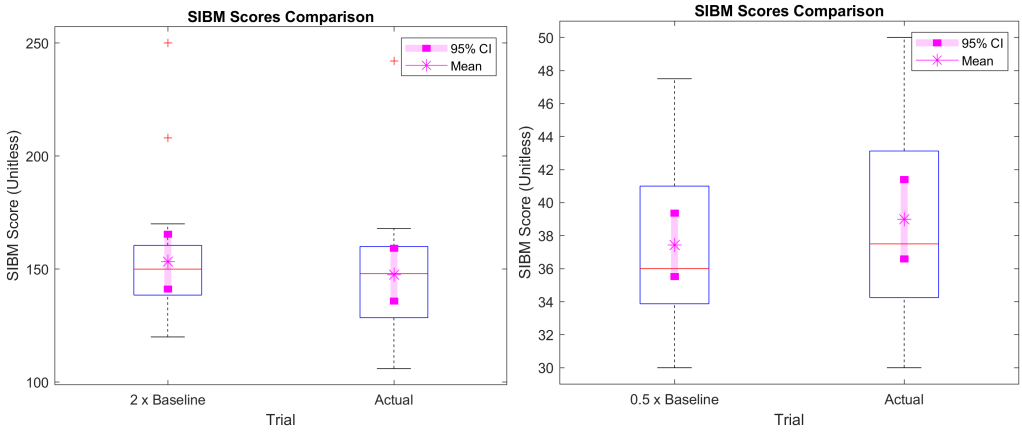
$$\begin{aligned}\text{SIBM} = 24.06 &+ 0.80 \times \text{FingerPressing} + 35.22 \times \text{Easy} - 12.19 \times \text{Hard} \\ &+ 0.35 \times \text{FingerPressing} \times \text{Easy} - 0.44 \times \text{FingerPressing} \times \text{Hard}\end{aligned}$$

Where Easy and Hard are mutually exclusive boolean values indicating the difficulty condition of the challenge. Table 17.8 summarizes the details, and lays out the equations for each condition. The Easy and Hard models match the regressions from Ch. 17.4.1 which we had deemed were reasonable. The Control model is slightly different than our baseline model[12], but they are similar considering the different data being tested and how reasonably well our baseline model performed in Ch. 17.3. The regression ANOVA (Table 17.8) shows both FP and Condition significantly contribute to the SIBM Score. There is also a significant interaction between FP and Condition ($F_{(2,67)} = 4.97$, $p = 0.01$), indicating the conditions are significantly different from the Control.

Is performance constantly scaled? Looking at the ratios of each condition's regression coefficients, using Control/Baseline as reference, we see Easy is generally twice as large (Control: 1.43, Baseline: 1.69) and Hard is about half as large (Control: 0.46, Baseline:

[10] $p = 0.019$
[11] $p = 0.028$
[12] SIBM = 26.70 + 0.68 × Finger Pressing

0.54). This leads us to think that the relationship between the participant's performance is just constantly scaled from the baseline with FP as we would expect. We investigate further by comparing the Easy and Hard conditions to constantly scaled baseline models.



(a) Easy (N=23) (b) Hard (N=25)

Figure 17.14: Comparing Predicted Scores to Actual Scores for SIBM.

Data Distributions: Participants perform relatively as expected in both conditions based on comparing predicted and actual scores (Fig. 17.14). The means and confidence intervals overlap in the predicted and actual values for both conditions. The median lines are visually similar between the predicted and actuals and they falls close to the respective means. The range of values is fairly similar as well. However, the distribution of differences between predicted and actual scores (Fig. 17.15) suggests the opposite. Only the Control condition's mean confidence intervals encompasses 0 (Easy: [-11.34, -0.14], Control: [-0.40, 6.08], Hard: [0.52, 2.61]). The Easy differences are largely variable; its mean and median imply the predicted values are overestimating participant performance. The Hard differences seem well-estimated and have a tighter distribution, but its mean and median imply our predictions underestimate performance.

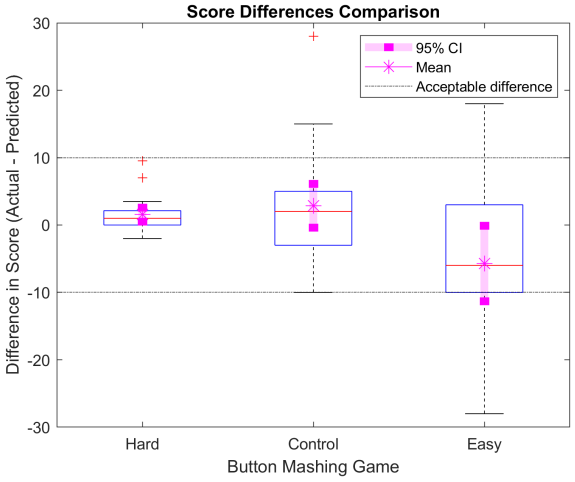


Figure 17.15: Score Differences for SIBM Across Conditions.

Easy vs. Upscaled Baseline. We multiply the baseline model by 2, and generate upscaled baseline predictions. We regress the Easy data and upscaled baseline predictions on FP with data source as a categorical variable. The regression ANOVA (Table 17.9) shows the Easy model is not significantly different from the upscaled baseline model, $F_{FingerPressing:Model(1,42)} = 0.91$, $p = 0.34$.

Looking at the Easy regression plot (Fig. 17.16) the slopes of the lines look different, but the scaled-baseline model seems to be relatively close to the calculated model. The two seem to diverge more as FP increases, making it seem like participants are underperforming in this condition.

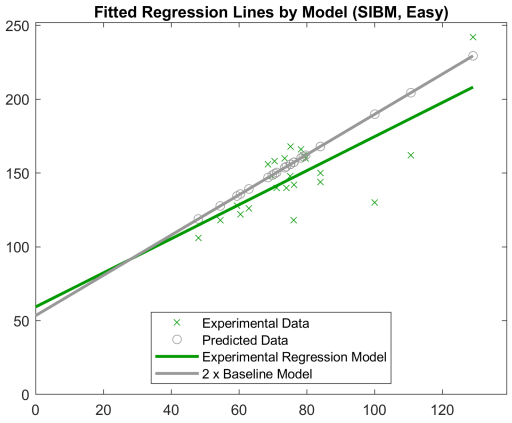


Figure 17.16: Easy vs. 2 x Baseline

Exp. Model	59.29 + 1.15 × Finger Pressing
2 x Baseline	53.39 + 1.36 × Finger Pressing

Variable	**Coeff.**	**t-stat**	**p-value**	**SS**	**DF**	**MS**	**F-Stat**	**p-value**
Intercept	59.29	4.87	< 0.0001	–	–	–	–	–
Finger Pressing	1.15	7.43	< 0.00001	21675	1	21675	131.42	< 0.0001
Model (Baseline)	-5.89	-0.34	0.73	1190.8	1	1190.8	7.22	0.01
FP x Model	0.21	0.96	0.34	150.82	1	150.82	0.91	0.34
Error	–	–	–	6927	42	164.93	–	–
Model: $R^2 = 0.94$, *Adj.* $R^2 = 0.93$, $F_{(5,67)} = 203$, $p < 0.00001$								

Table 17.9: Regression and ANOVA results showing effects of model type for Easy.

Hard vs. Downscaled Baseline. We scale the baseline model by 0.5 and generate downscaled predictions. We regress the Hard data and these downscaled predictions on FP with data source as a categorical variable. The Hard model is not significantly different from the down-scaled baseline model, $F_{FingerPressing:Model(1,46)} = 0.20$, $p = 0.661$ (see Table 17.10). Looking at the Hard plot, the lines almost overlap (Fig. 17.16). While the experimental regression seems to be under and overestimating its data points, it lines up with the down-scaled baseline almost exactly from about FP 50 to FP 80, which is where the average FP score lies. The lack of statistical difference between these regressions and a linear-scaling of the baseline model implies that differences we see in the data could be misleading because of human variability. However, it is also possible that our samples are just too small to detect the nuanced ways these models are different.

Figure 17.17: Hard vs. 0.5 x Baseline

Exp. Model	11.87 + 0.37 × Finger Pressing
0.5 x Baseline	13.35 + 0.34 × Finger Pressing

Variable	**Coeff.**	**t-stat**	**p-value**	**SS**	**DF**	**MS**	**F-Stat**	**p-value**
Intercept	11.87	3.74	0.0005	–	–	–	–	–
Finger Pressing	0.37	8.66	< 0.0001	935.03	1	935.03	139.46	< 0.0001
Model (Baseline)	1.478	0.33	0.74	2.83	1	2.83	0.42	0.52
FP x Model	-0.03	-0.44	0.66	1.31	1	1.31	0.20	0.66
Error	–	–	–	308.41	46	6.71	–	–
Model: $R^2 = 0.94$, Adj. $R^2 = 0.93$, $F_{(5,67)} = 203$, $p < 0.00001$								

Table 17.10: Regression and ANOVA results showing effects of model type for Hard.

17.5.2 MIBM

Model Performance

Easy. The model's performance is not great visually (poor $R^2 = 0.34$), with clusters of under performing participants (high FP, low score) implying a latent variable. The clusters could represent strategy, as six participants used two-hands (one finger on each key) to approach this challenge. As previously noted this could create a big difference in scores if participants were not paying attention to press the buttons at the same time. Notably in this group, three participants switched from using a two-handed strategy in the baseline to a one-handed strategy, so these participants may have felt this way and aimed to change their approach. It could also be a secondary ability at play, as mentioned in Ch. 16.

Hard. The values for the data are quite tight around the regression line. However, there are some clusters that could imply a latent variable. We do not think strategy is the underlying variable, as the dominant strategy was one-handed play (92% of participants in the group), with 3 participants switching from a two-handed to one-handed approach between the baseline and experimental conditions. It is more likely this represents another ability, as we proposed in Ch. 16. We believe that SelAtt may be this ability, as its correlation with score was unexpectedly large and positive, but not quite significant. It is possible that in this harder condition the use of cognitive abilities is more obvious and could be found more clearly with a larger sample size.

Comparing Regression Models

Are the conditions different? We regress the entire experimental dataset on FP with the condition as a categorical value. We use the Control condition as our reference point for interaction terms. This results in a significant model ($F_{(2,67)} = 115$, $p < 0.00001$, $R^2 = 0.90$), with the regression equation:

$$\begin{aligned}\text{MIBM} = {} & 6.89 + 0.96 \times \text{FingerPressing} + 55.98 \times \text{Easy} + 9.14 \times \text{Hard} \\ & - 0.01 \times \text{FingerPressing} \times \text{Easy} - 0.69 \times \text{FingerPressing} \times \text{Hard}\end{aligned}$$

Control	6.89 + 0.96 × Finger Pressing							
Easy	62.87 + 0.94 × Finger Pressing							
Hard	16.03 + 0.27 × Finger Pressing							
Variable	**Coeff.**	**t-stat**	**p-value**	**SS**	**DF**	**MS**	**F-Stat**	**p-value**
Intercept	6.89	0.34	0.74	–	–	–	–	–
Finger Pressing	0.96	3.54	0.001	7835.6	1	7835.6	35.55	< 0.0001
Condition	–	–	–	112450	2	56227	255.11	< 0.0001
(Easy)	55.98	2.26	0.03	–	–	–	–	–
(Hard)	9.14	0.33	0.74	–	–	–	–	–
FP x Condition	–	–	–	1241.9	2	620.93	2.82	0.07
(Easy)	-0.01	-0.04	0.97	–	–	–	–	–
(Hard)	-0.69	-1.89	0.06	–	–	–	–	–
Error	–	–	–	14767	67	220.4	–	–
Model: $R^2 = 0.90$, *Adj.* $R^2 = 0.89$, $F_{(3,67)} = 115$, $p < 0.0001$								

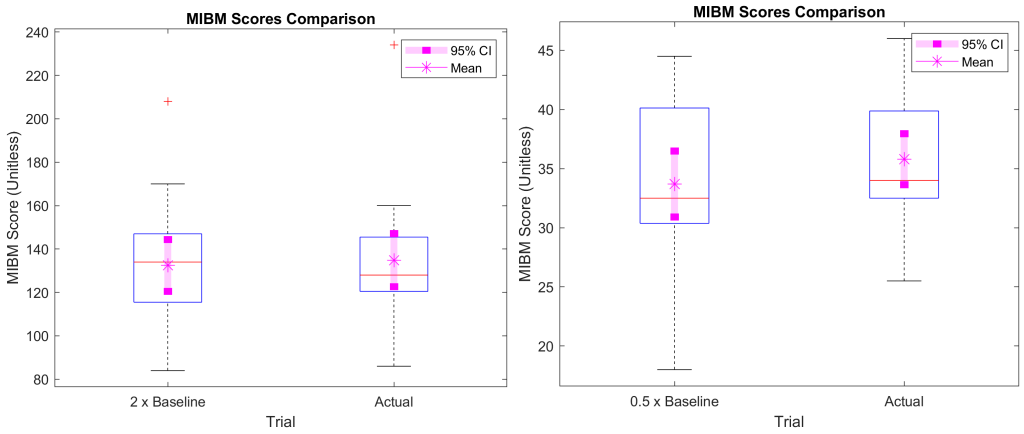
Table 17.11: Regression and ANOVA results showing effects of model type.

Where Easy and Hard are mutually exclusive boolean values indicating the difficulty condition of the challenge. Table 17.11 summarizes the details, and lays out the equations for each condition. The Easy and Hard models match the ones we found in Ch. 17.4.2. The Control model is quite different than our baseline[13], as we noted in Ch. 17.3. The ANOVA (Table 17.11) shows both FP and Condition significantly contribute to the MIBM Score. There is no significant interaction between FP and Condition ($F_{(2,67)} = 2.82$, $p = 0.07$), meaning the conditions may not be distinct enough from the control.

Is performance constantly scaled? The ratios of each condition's regression coefficients against the Control/Baseline do not align with our expectations. Easy's coefficient is 0.99 times the size of Control, but 1.7 times the size of Baseline. Hard's coefficient is 0.28 times the size of Control, but 0.48 times the size of Baseline. Considering the size of the intercepts and the odd fluctuations in the FP coefficients, this could be further evidence there is a latent predictor variable that is missing from the MIBM predictors. We further investigate this by checking the experimental regressions for the Easy and Hard conditions versus scaled baseline models.

Data Distribution: Participants performance seem to visually match our expectations in both conditions (Fig. 17.18). Their means and confidence intervals overlap between the distributions, and their medians are within their respective confidence intervals. However, the Hard condition's predicted median is just on the edge of the actual measure's bounding box (Fig. 17.18b), which could indicate the distributions are different.

[13] MIBM = 28.38 + 0.55 × Finger Pressing.



(a) Easy (N = 23) (b) Hard (N = 25)

Figure 17.18: Comparison of Predicted Scores to Actual Scores for Multiple Input Button Mashing Games.

Comparing the differences between actual and predicted scores, all the means are positive suggesting predictions are underperforming (Fig. 17.19). The confidence intervals for all conditions include 0 in their range, indicating the predicted and actual values' distributions are not significantly different (Easy: [-6.17, 11.04], Control: [-0.23, 7.75], Hard: [-0.32, 4.52]). However, the confidence intervals stretch further than their respective quantile box, indicating that perhaps there is a significant pull from outliers.

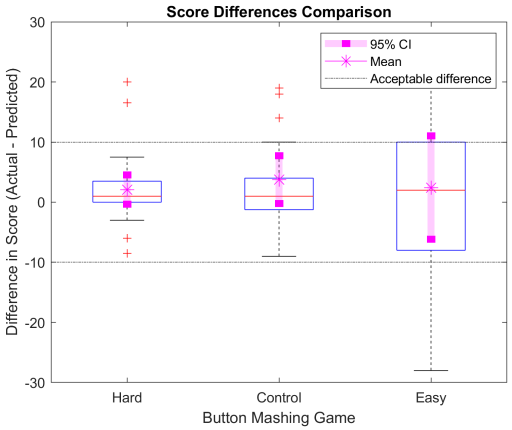


Figure 17.19: Score Differences for MIBM Across Conditions.

Easy vs. Upscaled Baseline. We multiply the baseline model by 2 and generate upscaled baseline predictions. We regress the Easy data and upscaled baseline predictions on FP with the data source as a categorical variable. The regression ANOVA (Table. 17.12) shows the models are not significantly different ($F_{FingerPressing:Model(1,42)} = 0.59$, $p = 0.56$). We see the regression lines intersect, but overall look close to each other (Fig. 17.20).

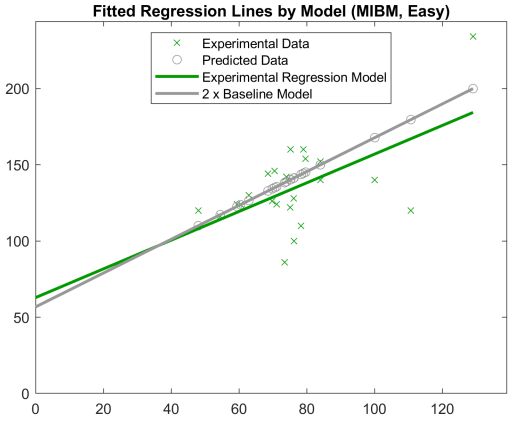


Figure 17.20: Easy vs. 2 x Baseline.

Easy	62.87 + 0.94 × Finger Pressing
2 x Baseline	56.75 + 1.11 × Finger Pressing

Variable	Coeff.	t-stat	p-value	SS	DF	MS	F-Stat	p-value
Intercept	62.87	3.97	0.0003	–	–	–	–	–
Finger Pressing	0.94	4.66	< 0.0001	14381	1	14381	51.65	< 0.0001
Model	-6.12	-0.27	0.79	529.19	1	529.19	1.90	0.175
FP x Model	0.17	0.59	0.56	97.20	1	97.20	0.35	0.56
Error	–	–	–	11694	42	278.42	–	–
Model: $R^2 = 0.56$, *Adj.* $R^2 = 0.53$, $F_{(2,42)} = 18$, $p < 0.0001$								

Table 17.12: Regression and ANOVA results showing effects of model type (Easy).

Hard vs. Downscaled Baseline. We scale the baseline model by 0.5 and generate downscaled predictions. We regress the Hard data and these downscaled predictions on FP with the data source as a categorical variable (details in Table 17.13). The models are not significantly different ($F_{FingerPressing:Model(1,46)} = 0.02$, $p = 0.89$). The regression lines begin to converge as FP increases (Fig. 17.21). The measured scores seem to differ from the predicted values more than we would expect.

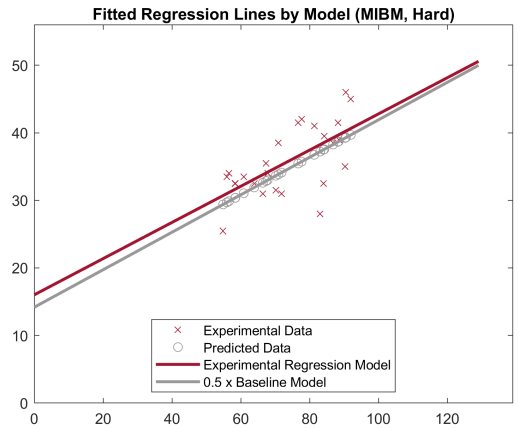


Figure 17.21: Hard vs. 0.5 x Baseline.

Exp. Model	16.03 + 0.27 × Finger Pressing
0.5 x Baseline	14.19 + 0.28 × Finger Pressing

Variable	Coeff.	t-stat	p-value	SS	DF	MS	F-Stat	p-value
Intercept	16.03	4.53	< 0.0001	–	–	–	–	–
Finger Pressing	0.27	5.66	< 0.0001	553.9	1	553.9	66.49	< 0.0001
Model	-1.84	-0.37	0.72	15.9	1	15.9	1.91	0.17
FP x Model	0.01[14]	0.14	0.89	0.17	1	0.17	0.02	0.89
Error	–	–	–	383.24	46	8.33	–	–
Model: $R^2 = 0.60$, *Adj.* $R^2 = 0.57$, $F_{(2,46)} = 22.8$, $p < 0.0001$								

Table 17.13: Regression and ANOVA results showing effects of model type (Hard).

[14]FP x Model = 0.0096218

17.5.3 AIBM

Model Performance

Easy. The model does not reasonably meet regression assumptions, likely due to the sample size being small. The primary issue is, in spite of our established theory (i.e. baseline models), SelAtt does not correlate to the scores. This could indicate that at the easy level, SelAtt is not as important enough to influence scores. Despite this, the model seems to perform well visually (though $R^2 = 0.32$ is poor) with some minor clustering that could indicate a latent variable. Strategy could be the variable as 12 participants (52%) used a one-handed strategy alternating between fingers, while 11 participants (48%) used a two-handed strategy with one finger on each key. It is possible that the choice of strategy affects the SelAtt used in this challenge, since it may require more robust SelAtt skills to maintain the pattern with two hands versus one. Comparing participant strategies to their baselines we see that 2 participants switched from a two-handed approach in the baseline to a one-handed approach in the experimental condition. While there is a good even split between strategies in this group, we would not be able to meaningfully regress using strategy as a predictor because each group falls so far below the minimum size threshold.

Hard. This model was deemed reasonable, and generally seems to perform reasonably well visually (but R^2=0.30). Minor violations of the regression assumptions seem to come from the small sample size. There is some clustering in the data, which could indicate a latent variable. While strategy is a possible option, the split between strategy is smaller in this condition (17 one-handed, 8 two-handed), leading us to believe it is less likely the culprit.

Comparing Regression Models

Are the conditions different? We regress the entire experimental dataset on FP and SelAtt using the condition as a categorical value (details in Table 17.14). This results in a significant model ($F_{(6,73)}= 69$, $p < 0.00001$, $R^2=0.90$), with the regression equation:

$$\begin{aligned}\text{AIBM} = {} & 11.11 + 0.50 \times \text{FingerPressing} + 4.20 \times \text{SelectiveAttention} \\ & + 24.98 \times \text{Easy} - 1.83 \times \text{Hard} \\ & + 0.18 \times \text{FingerPressing} \times \text{Easy} + 12.26 \times \text{SelectiveAttention} \times \text{Easy} \\ & - 0.28 \times \text{FingerPressing} \times \text{Hard} - 1.96 \times \text{SelectiveAttention} \times \text{Hard}\end{aligned}$$

Where Easy and Hard are mutually exclusive boolean values indicating the difficulty condition of the challenge. The Easy and Hard regressions match what we found in Ch. 17.4.3. The Control model looks quite different than our actual model from the baseline study[15], but matches what we found in Ch. 17.3 which was determined to be not statistically different. The regression ANOVA shows both FP and the Condition significantly contribute to the AIBM Score. SelAtt was not found to contribute ($F_{(1,73)}= 1.03$, $p= 0.31$). There is no significant interaction between the Condition and either FP ($F_{(2,73)}= 1.50$, $p= 0.24$) or

[15] AIBM = 20.75 + 0.28 × Finger Pressing + 11.38 × Selective Attention

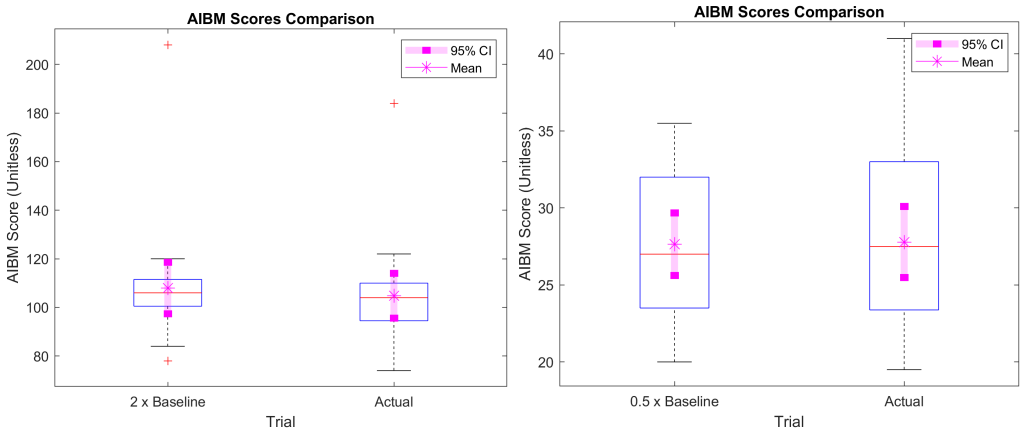
SelAtt ($F_{(2,73)}$= *0.42, p= 0.66*). This suggests the conditions may not be distinct enough from the control.

Control	11.11 + 0.50 × Finger Pressing + 4.20 × Selective Attention
Easy	36.09 + 0.68 × Finger Pressing + 16.45 × Selective Attention
Hard	9.28 + 0.22 × Finger Pressing + 2.23 × Selective Attention

Variable	**Coeff.**	**t-stat**	**p-value**	**SS**	**DF**	**MS**	**F-Stat**	**p-value**
Finger Pressing	0.50	2.07	0.04	3402.9	1	3402.9	24.28	< 0.0001
Selective Attention	4.20	0.33	0.74	144.31	1	144.31	1.03	0.31
Condition	–	–	–	68791	2	34396	245.43	< 0.0001
(Easy)	24.98	1.00	0.32	–	–	–	–	
(Hard)	-1.83	-0.08	0.94	–	–	–	–	
FP x Condition	–	–	–	419.49	2	209.74	1.50	0.23
(Easy)	0.18	0.62	0.54	–	–	–	–	–
(Hard)	-0.28	-0.86	0.39	–	–	–	–	–
SelAtt x Condition	–	–	–	117.07	2	28.54	0.42	0.66
(Easy)	12.26	0.69	0.49	–	–	–	–	–
(Hard)	-1.96	-0.12	0.90	–	–	–	–	–
Error	–	–	–	8969.3	64	140.15	–	–
Model R^2: 0.90, Adj. R^2: 0.88, $F_{(6,73)}$= 69, $p < 0.00001$								

Table 17.14: Regression and ANOVA results showing effects of model type.

Is performance scaled with the score modifier? Looking at the ratios of each condition's coefficients against Control/Baseline, Easy is larger across both FP (Control: 1.35, Baseline: 2.45) and SelAtt (Control: 3.92, Baseline: 1.45). Hard is smaller across both FP (Control: 0.43, Baseline: 0.79) and SelAtt (Control: 0.53, Baseline: 0.20). The Control and Baseline ratios for each condition seem quite different; this could be due to the small sample sizes leading to inaccurate experimental models. More interesting is the difference in scaling between FP and SelAtt is not the same. Looking at the Control ratios, SelAtt offers significant performance advantages in the Easy condition (despite not being well correlated), but scales more like FP in the Hard condition. In the Baseline ratios, SelAtt offers advantages in the Easy condition, but is significantly lower than we would expect in the Hard condition. The Hard condition's Baseline ratio for SelAtt suggests that players performance is severely punished for not being good at paying attention (i.e. having a low SelAtt measure to start). It is not clear from the ratios of the coefficients whether the conditions are simply constantly scaled from the baseline. We further investigate this by checking the experimental regressions for the Easy and Hard conditions versus scaled baseline models.



(a) Easy (N = 23) (b) Hard (N = 25)

Figure 17.22: Comparison of Predicted Scores to Actual Scores for Alternating Input Button Mashing Games.

Data Distributions: Participants performance between conditions is similar (Fig. 17.22). The means and confidence intervals overlap between the predicted and actual values. The medians are reasonably close to their respective means and between distributions. The differences distributions are well behaved across the conditions (Fig. 17.23). The Easy condition seems to be pulled downwards by its two low outliers, but outside of this skew seems to be quite reasonable. Each condition's mean confidence intervals include 0 (Easy: [-7.31, 0.87], Control: [-1.92, 0.88], Hard: [-0.62, 0.90]), which implies the predictions are not significantly different than the actual values.

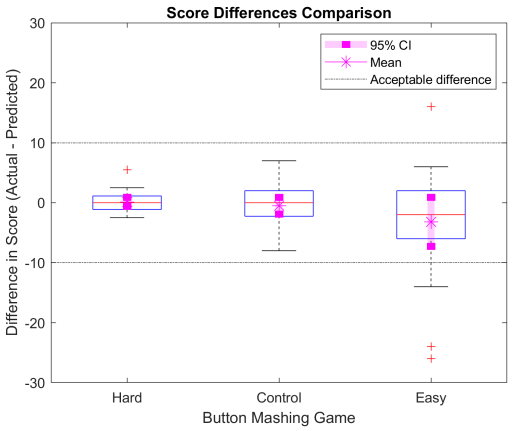
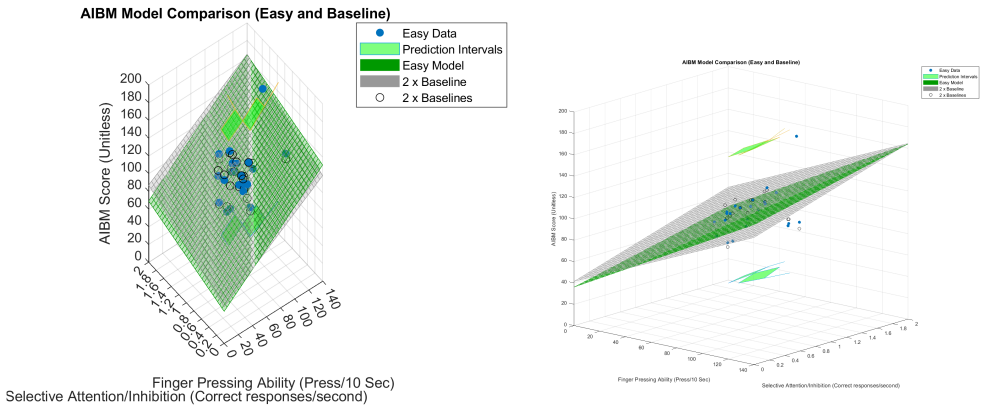


Figure 17.23: Score Differences for AIBM Across Conditions.



(a) (b)

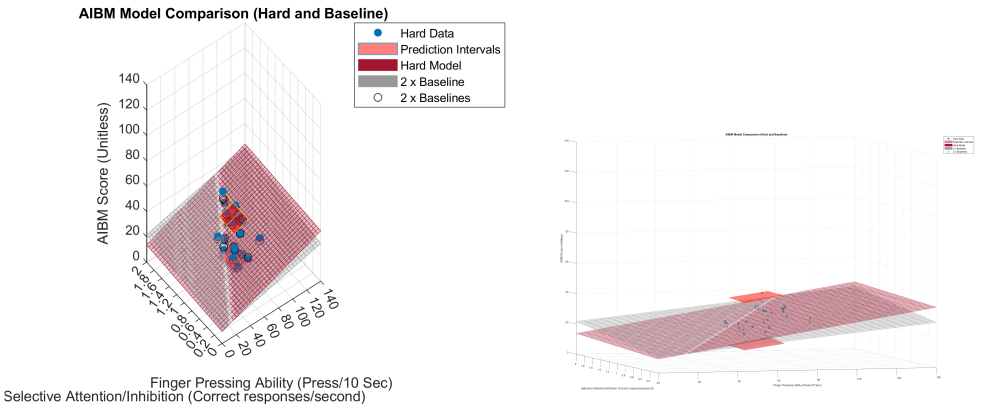
Figure 17.24: AIBM Regression Model for Easy Condition.

Easy vs. Upscaled Baseline. We scale the baseline model up by 2 and generate upscaled predictions. We regress the Easy data and upscaled predictions on FP and SelAtt

with the data source as a categorical variable. The ANOVA shows the Easy and upscaled baseline models are not significantly different (Table 17.15). There was no significant interaction between the data source and FP ($F_{(1,46)} = 0.30$, $p = 0.59$) or SelAtt ($F_{(1,46)} = 0.10$, $p = 0.75$). The Easy plot shows the planes that are relatively close to each other even though they are intersecting (Fig. 17.24).

Exp. Model	36.09 + 0.68 × Finger Pressing + 16.45 × SelAtt				
2 x Baseline	41.50 + 0.55 × Finger Pressing + 22.76 × SelAtt				
Variable	**SS**	**DF**	**MS**	**F-Stat**	**p-value**
Finger Pressing	176.03	1	176.03	29.60	< 0.0001
Selective Attention	43.43	1	43.43	3.94	0.05
Model (Baseline)	12.98	1	12.98	0.37	0.55
FP x Model	8.74	1	8.74	0.30	0.59
SelAtt x Model	8.26	1	8.26	0.10	0.75
Error	522.21	44	11.87		

Table 17.15: ANOVA results showing effects of model type (Easy vs. Baseline).



(a) (b)

Figure 17.25: AIBM Regression Model for Hard Condition.

Hard vs. Downscaled Baseline. We scale the baseline model by 0.5 and generate downscaled predictions. We regress the Hard data and downscaled predictions on FP and SelAtt using data source as a categorical variable. The ANOVA shows the hard and downscaled baseline models are not significantly different (Table 17.16). There was no significant interaction between the model type and FP ($F_{(1,50)} = 0.74$, $p = 0.40$) or SelAtt ($F_{(1,50)} = 0.70$, $p = 0.41$). The Hard plots (Fig. 17.25) show the actual performance seems to be better than the predicted performance. Looking at where the planes intersect we get a sense that the effect of SelAtt is not being well displayed here, since that seems to be the variable that misaligns these planes.

Exp. Model	9.28 + 0.22 × Finger Pressing + 2.23 × SelAtt
0.5 x Baseline	10.37 + 0.14 × Finger Pressing + 5.69 × SelAtt

Variable	**SS**	**DF**	**MS**	**F-Stat**	**p-value**
Finger Pressing	176.03	1	176.03	14.83	0.0004
Selective Attention	43.43	1	43.43	3.66	0.006
Model (Baseline)	12.98	1	12.98	1.09	0.30
FP x Model	8.74	1	8.74	0.74	0.40
SelAtt x Model	8.26	1	8.26	0.70	0.41
Error	522.21	44	11.87	–	–

Table 17.16: ANOVA results showing effects of model type (Hard vs. Baseline).

17.6 Relationship between Challenges and Loads

After exploring the relationships between the experimental models and their relationships with baseline models, we synthesize this information to assess our hypotheses directly (Table 17.17). **Overall, performance scales as expected for all conditions; competency profile loads change in the right direction, but not as much as we expected.** The following sections focus on explicitly explaining the performance-score modifier relationship (H1) and the competency profile changes (H2, H3).

	SIBM		**MIBM**		**AIBM**	
	Overload	**Underload**	**Overload**	**Underload**	**Overload**	**Underload**
H1	✔	✔	✔	✔	✔	✔
H2	✖	✖	✖	✖	–	–
H3	–	–	–	–	✖	✖

Table 17.17: Hypotheses Results Comparing Under and Overload models to Baselines.

17.6.1 SIBM

Performance and Score Modifiers: From our investigations, it seems like the **SIBM performance scales in step with the score modifier**. Fig. 17.26 shows the models, their confidence and prediction intervals, and experimental data for each condition. We include reference lines for the average FP rate for the entire participant sample, and one standard deviation around it. Easy and Baseline's prediction intervals overlap left of FP = 80. This is likely due

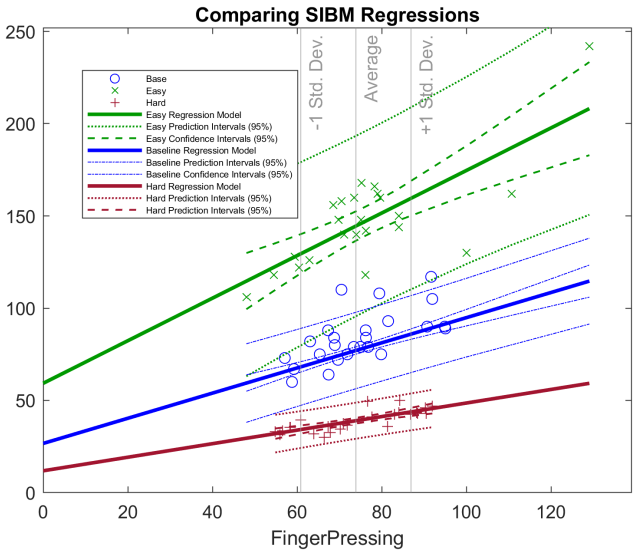


Figure 17.26: Comparison of all SIBM Regression Models vs. Experimental Data

to Easy's large uncertainty because of the difficulty scaling amplifying any noise in the data. We see the high performing Control participants are in Easy's prediction intervals, suggesting the low Easy participants are significantly under performing and could be doing so because they are not "trying their best". It seems when tuned trivially easy, **players may reserve their energy and purposefully under perform knowing they are safely able to complete the challenge**. In comparison, it seems like participants in the **overloaded conditions are more consistent in their performance** because they are giving it their all.

Competency Profiles and Score Modifiers: We calculate the standardized coefficients for the Easy (0.56) and Hard (0.82) model to compare against the Baseline (0.74). We graph the Underload (i.e. Easy), Baseline, and Overload (i.e. Hard) competency profiles (Fig. 17.27). We see that even though the performance was linearly scaled, Hard's FP load is only 1.12 times the size of Baseline, and Easy is only about 0.76 times the size Baseline. While there was load scaling in the expected directions, these do not strictly meet our hypotheses. Using these points we approximate the relationship between changes in the score modifier and FP load as: Finger Pressing $= 0.91 - 0.18 \times$ Score Modifier.

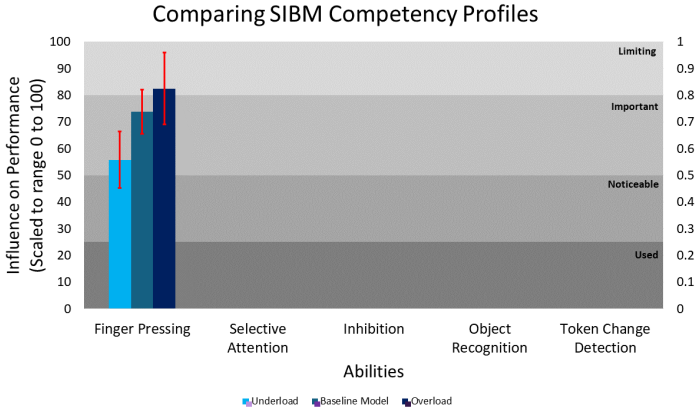


Figure 17.27: Comparing SIBM competency profiles at different loads.

17.6.2 MIBM

Performance and Score Modifiers: While our investigation shows that **performance scales with the score modifier**, we have to keep in mind that the **models require more investigation**. We plot the Easy and Hard models alongside the Baseline (Fig. 17.28). We include the experimental data, the models' confidence and predictions intervals, and reference lines indicating the average FP rate ± 1 standard deviation. We see Easy's lower data points fall inside the Baseline prediction intervals and fit in

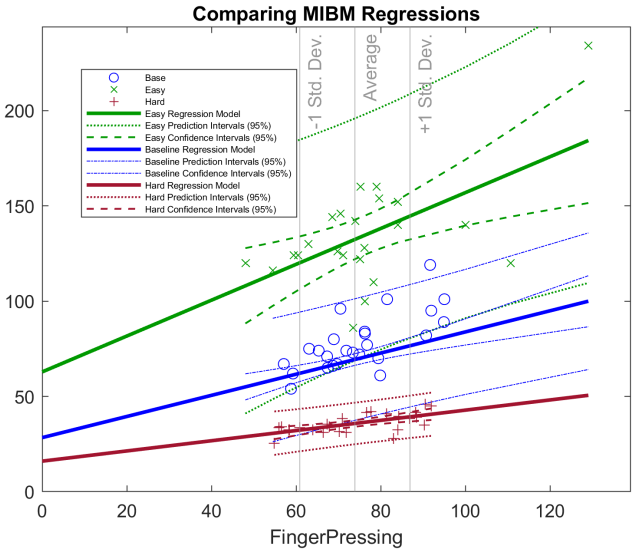


Figure 17.28: Comparison of all MIBM Regression Models vs. Experimental Data

among the Control values. This could be happening because of underperforming Easy participants. These **"under performing" outliers could be excessively impacting the group means**, causing them to look similar. This **may explain why the interaction term did not show a significant difference between conditions** (see 17.5.2). The Baseline predictions also overlap the Hard confidence intervals, but the data points do not intermingle. This leads us to believe that the Hard is significantly different from the Baseline, which in turn may be what is pulling the interaction term from our comparison ANOVA towards being significant.

Competency Profiles and Score Modifiers: We calculate the standardized coefficients for the Easy (0.43) and Hard (0.67) models to compare to the Baseline (0.50). We graph the Underload (i.e. Easy), Baseline, and Overload (i.e. Hard) competency profiles (Fig. 17.29). Hard's FP load is about 1.34 times Baseline, and Easy is 0.86 times Baseline. While the increase is happening in the expected direction, the magnitude does not match our hypothesis. Using these data points we approximate the relationship between changes in the score modifier and FP load as: Finger Pressing $= 0.71 - 0.15 \times$ Score Modifier.

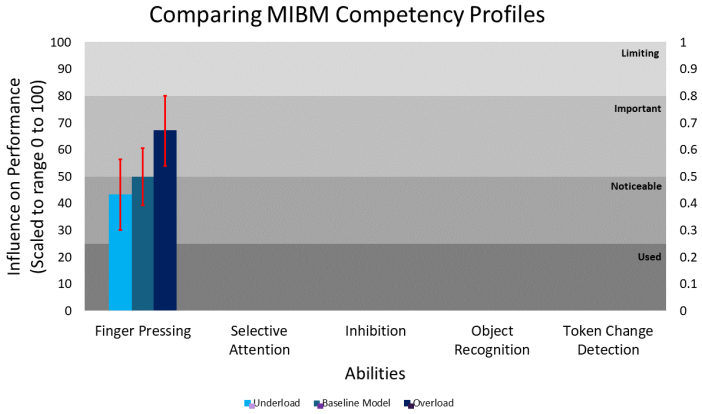
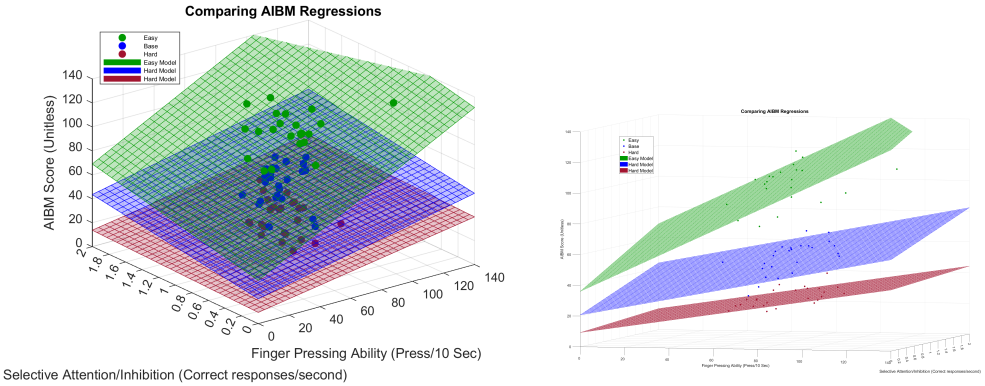


Figure 17.29: Comparing MIBM competency profiles at different loads.

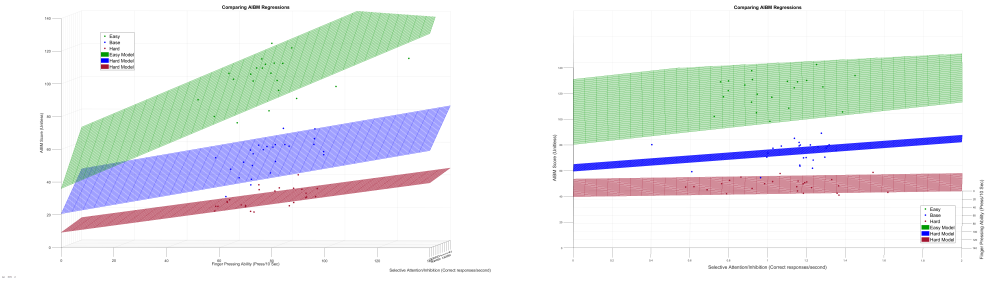
17.6.3 AIBM



(a) (b)

Figure 17.30: Comparison of all AIBM Regression Models vs. Experimental Data

Performance and Score Modifiers: From our investigations, **performance seems to be tied to score modifier**. We plot the regression planes, and the experimental data side by side (Fig.17.30). For readability purposes we leave out the prediction and confidence intervals, and reference lines for the ability measures. While for the most part data points are close to their respective planes, there are spaces where they seem to butt up against the prediction intervals for its adjacent condition. The most interesting element of comparing performance is seeing how SelAtt (the secondary ability) changes importance between conditions (Fig.17.31b), particularly compared to FP changes (Fig. 17.31a). We previously discuss this in Sec. 17.5.3, but **it seems that SelAtt is a "bonus" skill in Easy (improving scores but a high value is not a success requirement) but becomes more required as difficulty increases**. Recall that SelAtt did not correlate to the Easy scores (which made the regression unreasonable); SelAtt as a "bonus" skill in trivial difficulties could explain why does not correlate for just this condition.



(a) Finger Pressing Axis (b) Selective Attention Axis

Figure 17.31: AIBM Regression Models rotated to highlight specific growth along axes.

Competency Profiles and Score Modifiers: We calculate the standardized coefficients for the Easy (FP: 0.36, SelAtt: 0.16) and Hard (FP: 0.51, SelAtt: 0.09) models to compare to the Baseline (FP: 0.38, SelAtt: 0.28). We graph the Underload (i.e. Easy), Baseline, and Overload (i.e. Hard) competency profiles (Fig. 17.32). For FP load, Hard is about 1.33 times Baseline, and Easy is 0.95 times Baseline. While the changes are happening in the expected directions, the magnitudes do not match our hypotheses. We also look at SelAtt load; Hard is 0.33 times the size of Baseline, and Easy is 0.56 times Baseline. It is unclear why the changes in SelAtt are like this; however, we did expect this in our hypothesis. Using these data points we approximate the relationship between changes in the score modifier and FP and SelAtt loads as:

$$\text{Finger Pressing} = 0.52 - 0.09 \times \text{Score Modifier}$$

$$\text{Selective Attention} = -0.33 \times \text{Score Modifier}^2 + 0.87 \times \text{Score Modifier} - 0.26$$

17.7 Conclusion

This chapter is an exploratory look at the relationship between a challenge's ability loadings (via competency profiles) and mechanical achievability (via performance) for the purpose of answering RQ2.

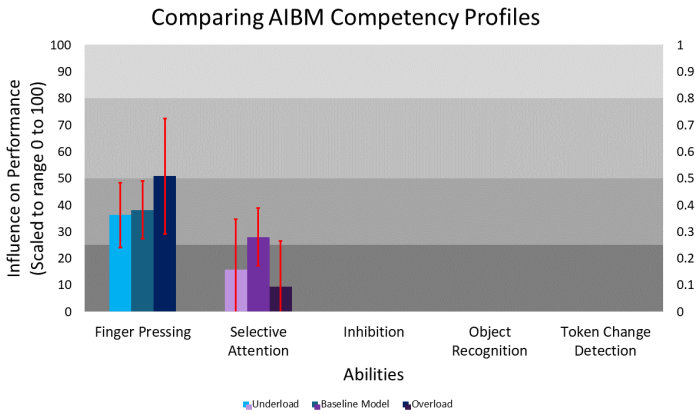


Figure 17.32: Comparing AIBM competency profiles at different loads.

Verifying Baseline Models: We begin by verifying our baseline models of button mashing challenges against new data from the same participants. The data analysis suggests that participants were performing within human variability error between button mashing instances. As well, results seem to imply that the baseline and control models were statistically identical (i.e. for each challenge there was no significant interaction effect between performance and whether the data was from the Baseline or Control set).

Analysing the Relationship Between Performance and Finger Pressing Load: We analyse player performance at button mashing challenges tuned to different difficulties, and compare the performance models for the Underloaded (i.e. Easy), and Overloaded (i.e. Hard) instances to the Control (i.e. Baseline) models. Our results highlight the changes in a player's button mashing challenge performance as a result of FP load. Specifically, we observe that:

1. Performance in SIBM seems solely influenced by FP load;
2. Performance in MIBM seems largely influenced by FP load at Easy and Baseline tunings, but there seems to be a latent variable (potentially SelAtt given the correlations) that exerts influence at the Hard tuning; and,
3. Performance in AIBM seems primarily influenced by FP load, and secondarily by SelAtt.

Quantifying the Mechanical Difficulty Relationship Between Abilities and Score Modifier: We use these performance models to provide preliminary insight into the mechanical difficulty relationship between the tuning of score modifiers to the amount of FP load in the challenge (represented in the competency profile). Specifically these results suggest:

1. SIBM's FP is linearly related to the score modifier ($\text{FP} = 0.91 - 0.18 \times \text{Score Modifier}$);
2. MIBM's FP load is linearly related to the score modifier ($\text{FP} = 0.71 - 0.15 \times \text{Score Modifier}$);
3. AIBM's FP load is linearly related to the score modifier ($\text{FP} = 0.52 - 0.09 \times \text{Score Modifier}$); and,

4. AIBM's SelAtt load is quadratically related to the score modifier (SelAtt $= -0.33 \times$ Score Modifier$^2 + 0.87 \times$ Score Modifier $- 0.26$)

By establishing a preliminary understanding of the performance relationship between FP load and button mashing challenges, and the mechanical difficulty relationship between abilities and score modifier, we have a starting answer to RQ2.

Limitations in Results: These results are limited by factors discussed in Ch. 15.5, particularly the small sample size, uncertainty in the measures, and noise in the data.

Future work. A more thorough investigation of AIBM should be the first step in furthering this work. AIBM shows interesting and unclear effects on the secondary ability (Selective Attention). From the current data, AIBM seems to change competency profile shapes with difficulty as its secondary ability becomes more or less important. We suspect something similar is happening with MIBM, where a secondary variable like Selective Attention becomes more necessary at higher difficulties. To confirm this we would need more targeted studies with larger sample sizes. Another direction for future work is to explore dynamically adjusted condition to try and control participant effort modulation which could be causing data overlap.

Take home points

From this chapter we learned the following meta-lessons:

- Competency profiles represent ability load for challenge tuning.
- The baseline models for Single Input Button Mashing and Alternating Input Button Mashing are reasonably good at fitting new data.
- The regression issues with Multiple Input Button Mashing (particularly the Overload condition) adds more evidence to a latent variable in the competency profile.
- Performance in button mashing challenges seem directly tied to the limiting ability and scales with score modifier as expected.
- Player's effort at a challenge may be directly tied to how difficult they perceive it to be.
- Button mashing competency profiles do not change finger pressing load one-to-one with score modifier (i.e. what "counts" as a button press).
- Competency profiles may change depending on difficulty tuning (as with AIBM Underload and MIBM Overload).

Chapter 18

Mechanical Experience of Competency Profiles (Study 1C))

We conduct a convergent mixed methods research [101] study to explore the player's perceptions of the button mashing challenges in their experimental conditions. This serves to address RQ3, as we explore the relationship between controlled mechanical experiences (facilitated by mechanical difficulty changes) to the larger holistic player experience.

We start by constructing our theoretical framework. We focus on concepts and theories which we rely on when contextualizing and analysing the data. The theoretical framework replaces a specific hypothesis as it summarizes what we expect from the data and how we will be interpreting subjective elements. We then explain the study design and summarize participant details. We provide a short researcher reflexivity statement to further contextualize our interpretations of the data, and increase integrity and transparency. We then spend a moment examining participants to understand how their intersecting identities and positions affect their responses. We then present the data in two separate analyses: first the quantitative data, then the qualitative. For each type of data, we explain our data analysis approach, present the results, and discuss our interpretations of the data on its own before coming together at the end for a meta-discussion about how the results support each other and interact. We end this chapter by expressing how these results will impact our design of "jutsu".

A Quick Aside...

This study has a lot of data. For readability, we present samples and summaries of data in the body of the thesis. Full data and results are in App. H.

18.1 Theoretical and Conceptual Framework

We take a social constructionist perspective to this research meaning that we see reality as constructed through the social discourses and practices of people [86]. We establish some basic theoretical foundations of how experiences are constructed and what other social constructs influence these experiences. While the whole thesis can be thought of as developing this

larger conceptual framework for understanding player experiences, players, and games, we particularly highlight a handful of immediately relevant concepts.

Player Experiences. Player experiences (PX) are constructed phenomena composed of many experiential types (see Experiential Tetrad - Fig. 3.1). These experiences exists at multiple levels of abstraction and across different times, as reflected in Nacke and Drachen's PX framework (Fig. 18.1) [315]. Since we focus on mechanical experiences, **we anticipate discussions to centre the technical experience and individual experience**. Leveraging these conceptual frameworks further, we can **anticipate that an individual player's *game literacy*, and *gamer identity* play a significant role** in how they experience our difficulty-adjusted games.

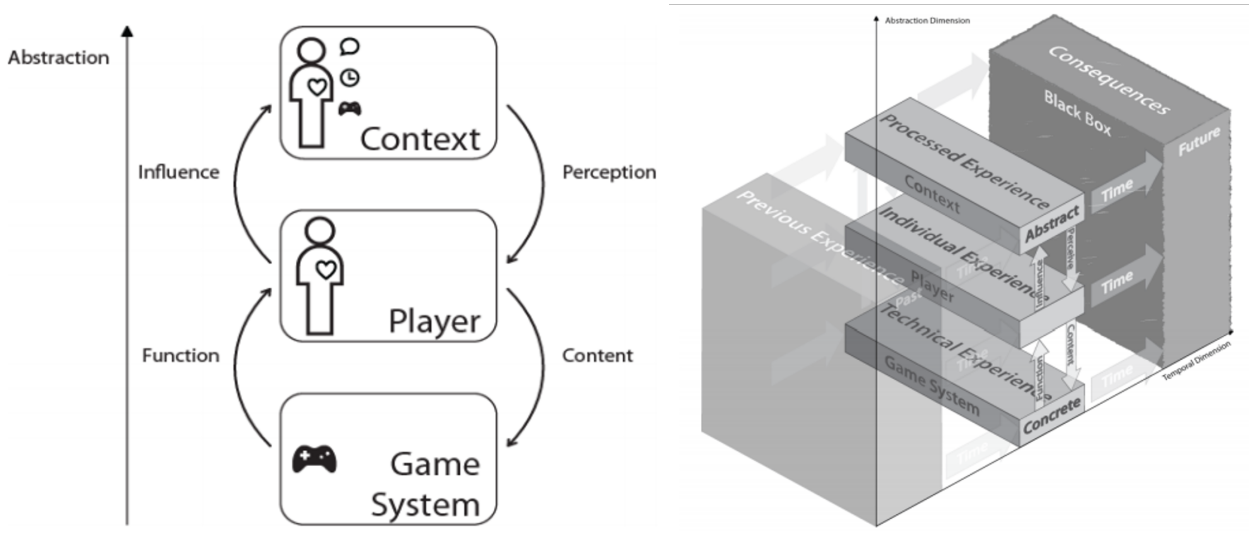


Figure 18.1: Nacke and Drachens Player Experience Framework reproduced from [315].

Game Literacy. Game literacy is broadly the ability to understand and make meaning in games at different levels of abstraction by employing various skills and competencies built through play, design, and analysis [14, 540, 542]. For the average player, their gaming literacy is **developed through their personal gaming experiences, interaction with gaming paratexts, and the larger gaming culture**. In this way players develop a wealth of knowledge about common mechanics, goals, interactions, and optimal strategies which improve their performance in games. This wealth of experience can **impact a player's perceived usability, competence, immersion, and positive affect in similar games** (e.g. games in the same franchise 206, or genre).

Gamer Identity. Self-categorization theory [484] explores how social identity groups both *describe* and *prescribe* their members' attitudes and behaviours. Identities are therefore performative (built through ongoing adherence to attitudes and behaviours) and precarious (under scrutiny by others who validate that identity) [72]. An **individual's performance of their identity is constantly compared against the prototypical group member**.

Shaw [432] summarizes "gamer" identity as being performed through acts of *consumption* (e.g. time commitment, economic investment, subcultural capital) and *dedication* (e.g. social capital). However, this only accounts for some of the behavioural elements of the prototype — to fully understand gamer identity we need to combine this knowledge with the "gamer" stereotype. For "gamers" the traditional stereotype paints a picture of a young White man [106, 359, 431, 523], who is commonly heterosexual [431], unattractive [243], socially inept and unmotivated for real world rewards [243, 452] but likely has some technical expertise [112]. When it comes to play, the stereotype is that "real gamers" want challenging competitive experiences [94, 359], and therefore focus on a subset of "real games" and playstyles [145, 223, 286]. This is reinforced through the identity sub-labels and rhetorics found in online gaming communities which create a hierarchy of "realness" [60, 112]. While continually disproven both academically (e.g. 170, 243, 509, 523) and through gamer census data (e.g. 15, 16), the stereotypes still persist even among self-identified gamers who lobby it at others or use it as a reason to discredit themselves as gamers [431, 452]. The persistence of the "gamer" stereotype in spite of contradicting evidence, implies that it *is* the prototypical view of a gamer even inside the gaming community. **Tension with this stereotype leaves individuals unsure of their "gamer" identity even when they exhibit stereotypical behaviours** (e.g. 112, 359, 432).

Identity Threat. When a precarious identity is challenged either culturally or individually strongly identifying group members may experience identity threat [37, 57], leading members to **reassert themselves by performing more prototypical group behaviours**, and denigrating outgroups [57]. Self-Affirmation Theory[1] [448] outlines three typical responses to identity threat: *accommodation* (accepting the threat and changing themselves), *amelioration* (handling the threat by reframing, dismissing, denying, or avoiding it), or *self-affirmation* (reflecting on aspects of our self-image unrelated to the threat) [433]. For gamers, under the prototypical belief of a gaming meritocracy [220, 365], failure can result in identity threat inciting gamers to **self-regulate individually through defensive biases** (e.g. attributing the loss to the game or circumstance [220, p. 17-19]; self-defeating behaviours like expecting negative outcomes or not recognizing poor performance [510]), and **self-affirm culturally through reinforcing the stereotype** (e.g. stating women are inferior players due to biological and cultural reasons [296]), harassing perceived outsiders (e.g. being hostile towards women gamers [226]; toxic practices [60, 365]), **or reframing the threat** (e.g. attributing negative outcomes to external prejudice and having a more positive opinion of being a gamer [508]).

Meaning for our study. To understand latent meanings and implications of responses we must consider:

- Responses will likely focus on the concrete levels of experience (i.e. technical and individual);

[1] This overlaps with Attribution Theory [141] which explores how people connect events to causes. Juul [220] notes that gamers can either attribute failure to themself (i.e. accommodation), the game (i.e. amelioration) or circumstance (i.e. amelioration).

- Gaming literacy and gamer identity will influence what they perceive, how they feel about it, and the way they communicate;
- Identity threat may influence responses leading to prototypical gamer behaviours/responses.

Therefore we must familiarize ourselves with contemporary gamer culture, and look at how our participants fit into it.

18.2 Study Design

We conduct a convergent mixed-methods[2] study [101], using a questionnaire variant [100], to understand PX for button mashing challenges at different difficulties. We collect Player eXperience Inventory (PXI) responses as quantitative data, and open-ended response questions as qualitative data. Using a questionnaire variant means we will not have rigorous qualitative data since the open-ended questions are just appended to the quantitative survey, and we cannot follow up with participants as we would in an interview. While this means the data would not hold up on its own as a qualitative study, it can still produce emergent themes to enhance and discuss the quantitative data set [100]. Details of the study design, procedures, and apparatus are in Ch. 15.

18.2.1 Researcher Reflexivity

Sasha has been academically researching games since 2016, which affects how she interprets participant responses due to her awareness of game studies and human-computer interaction work. Sasha is an avid game player for over 25 years, covering a variety of genres (e.g. shooters, hack-and-slash, puzzle, party, visual novel, point-and-click adventure, roleplaying, strategy, platformers, etc). This influences the approach to the overall thesis including the focus on creating models of games at the level of atomic challenges. Sasha identifies as a gamer with previous inclinations to being a "hardcore" gamer (see Brett and Soraine [60] for meanings of different sub-labels) and she participates in broader gamer and gaming culture. This gives Sasha a partial insider perspective as her gender (not cis-male) and racial (not White) identity do not align with the "gamer" stereotype. While the overlapping gamer identities allow for more nuanced interpretation of the responses and insight into latent meanings, we do acknowledge the potential bias of over-relying on perceived group stereotypes because of this. We try to offset this through relying on our conceptual framework to support insights and meanings we find.

[2]Details on convergent mixed method research are in App. H.1.

18.2.2 Participants

75[3] participants complete our study (Men: 46, Women: 25, Queer[4]: 4). We collect participants' ages, genders, gamer identity, gaming history, and gaming habits (see Table 18.1 for sample). Based on the average responses, a representative participant is a 21.9 year-old (σ 4.2, range: 18 to 37) man who self-identifies as a gamer (68%: 35 men, 12 women, 4 queer), has been playing games for 10+ years (60%: 32 men, 9 women, 4 queer), and averages 5+ gaming sessions a week (44%: 24 men, 7 women, 2 queer) at 1 to 3 hours per session (48%: 24 men, 12 women, 0 queer) on his computer (85%). We also collect their gaming platforms and a sample of games they play (see Table 18.2 for sample).

	Demographics			**Gaming Information**		
ID	**Age**	**Gender**	**Gamer**	**History**	**Length (hours)**	**Sessions per week**
P01	30	Man	Yes	10+ years	0.5 to 1	5+
P02	21	Man	Yes	10+ years	1 to 3	5+
P03	28	Man	Yes	10+ years	1 to 3	2 to 4
.	.	.	.	.	.	.
P18	19	Woman	No†	10+ years	1 to 3	Irregular
P19	19	Man	No†	5 to 9 years	0.5 to 1	1
P20	26	Man	No	10+ years	1 to 3	5+
P21	21	Woman	Yes†	10+ years	1 to 3	2 to 4
P22	30	Queer	Yes†	10+ years	3+	5+
.	.	.	.	.	.	.
P73	29	Man	Yes	10+ yeas	1 to 3	5+
P74	20	Man	Yes	10+ years	3+	5+
P75	33	Woman	Yes	10+ years	1 to 3	2 to 4

†: indicates a more complicated response than Yes/No

Table 18.1: Sample of Participant Demographics (11/75 rows). Full information in Table H.10.

18.3 Analysing Participants

We take a critical look at the participants' gamer identity, history, and habits. We aim to get a better understanding of how these intersect in ways that could help us interpret responses more accurately.

[3]This study does not need to remove any participants as we have complete data from their questionnaires (the primary data source), as well as their ranks in the baseline, experimental conditions (secondary data sources).

[4]We include specific identities in the Participant Information table (Table 18.1), but for simplicity in numerical reporting we categorize Trans, Non-Binary, Queer and fluid or unshared identities as "Queer".

ID	PC	Ph/T	Con	HCon.	Sample of Reported Games/Genres
P01	✓	×	✓	×	LoL, Zelda, CoD, Stardew Valley, Mario Party
P02	✓	✓	✓	×	FIFA, CoD, Fortnite
P03	×	×	✓	×	Last of Us, GTA V, God of War
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
P18	×	×	✓	×	CoD
P19	✓	×	✓	✓	Minecraft, Kingdom Hearts, Forza Horizon, CS:GO
P20	×	×	✓	×	FIFA, Overwatch, Warzone, Spiderman
P21	×	✓	✓	×	Animal Crossing, Mario Kart, Mario Party, Overcooked 2, Cookie Run Kingdom
P22	✓	×	✓	✓	MMORPGs, MOBAs, party games, fighting games, Mario Kart
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
P73	✓	✓	×	✓	Palworld, LoL, Pokemon
P74	✓	✓	✓	✓	Action-Adventure, JRPGs, Puzzle, Shooters, Platformers, Co-op, Arcade
P75	✓	×	✓	×	Mario Kart, FF XII, Layers of Fear, Love Nikki, Pokemon Snap, Shadow of the Colossus
Legend	Computer (PC), Smartphone/Tablet (Ph/T), Console (Con), Handheld consoles (HCon), League of Legends (LoL), Call of Duty (CoD), Grand Theft Auto (GTA), Counterstrike: Global Offensive (CS:GO), Legend of Zelda (Zelda), Final Fantasy (FF), Japanese Roleplaying Game genre (JRPG), Massively Multiplayer Online Roleplaying Game genre (MMORPG), Multiplayer Online Battle Arena genre (MOBA)				

Table 18.2: Sample of Participant Gaming Context (11/75 rows). Full information in Table H.12.

Uncertain Gamers. Eight participants express uncertainty about their gamer status. Those who seem positively-inclined to the identity (P21, P22, P36) have long gaming histories (10+ years) and frequent weekly sessions (2 to 4) which generally fit or exceed the average session length. The negatively-inclined ones (P14, P18, P19, P34, P57) report long gaming histories (5 to 9 years: P19, P34, or 10+ years: P18, P57), irregular gaming sessions (P14, P18), or shorter average sessions (0.5 to 1 hour: P19, P34, P57). **Uncertainty implies a complex relationship to the gaming stereotype, not inherently gaming experience or knowledge.**

Non-gamers. 75% report gaming for at least one year (29% reporting more than 10 years), with 37.5% report between one to four gaming sessions a week — numbers that hardly seem like low amount of play. While "non-gamers" report shorter sessions than the average participant, there is evidence that self-report measures are biased by social and cognitive processes (e.g. 362) and that video game playtime is generally underestimated by players [45] as time is perceived to pass faster in flow states [402]. This makes it seem like **self-identification as a non-gamer reflects the participants comparing themselves to the gamer stereotype, rather than having significant inherent differences in their relationship to games**.

"Gamer" Habits. Patterns aligning with the gamer stereotype emerge in platform preferences (Computer: 85%, Smartphone/Tablet: 57%), common games (e.g. Valorant, League of Legends, Minecraft, FIFA) and genres (e.g. Action, Action-Adventure, MOBA, FPS, Sports). Insight into how these preferences are related to gamer identity and stereotype will allow for more nuanced response interpretations.

Platform. Computer gaming carries social gaming capital [60, 94], making it unsurprising it is the most common platform. Our gamer participants are likely very familiar with manipulating games through keyboard (95% report being familiar and confident with a keyboard). In gaming culture, mechanical keyboards are discussed as necessary for optimal performance both in community posts (e.g. Reddit [501]), and scholarly work on gaming performance (e.g. [280, 298]).

Games and Strategy. The most commonly listed genres (e.g. Action, MOBA) and games (e.g. Valorant, League of Legends) indicate that many of our participants exist in *competitive gaming culture* [60]. This competitive space embodies the gaming meritocracy mindset and so considers all competition as inherently strategic. To competitive-oriented gamers "button mashing" is a reviled strategy for new players and non-competitives (e.g. 531).

Games and Age. In light of participant ages and gaming histories, the commonly listed games imply that our **our average participant may not be exposed to button mashing as a challenge**. Button mashing as challenge became heavily tied to quick-time events (QTEs), which larger gaming culture began to openly detest around the 2010s (e.g. 292). This led to button mashing challenges being used less frequently in large and competetive games, though it still exists in party and mini-games.

What we learn about the participants:

- Gamer identity or lack-thereof does not indicate lack of gaming experience or poor game literacy.
- Specific hardware (e.g. mechanical keyboards) and strategy discussions can be indicators of prototypical gamer views.
- Participants may not be familiar with button mashing as a challenge, and so interpret it as button mashing as a strategy (which can trigger identity issues on failure).

18.4 Quantitative Strand: PXI Responses

18.4.1 Analysis Method

We analyse the PXI data in five steps (Fig. 18.2). We start by **recoding participant item responses from the original 1 (strongly disagree) to 7 (strongly agree) into a range from -3 (strongly disagree) to +3 (strongly agree)**. This way the response's sign represents its valence (positive or negative), and its value indicates intensity, and scores of 0 are neutral (neither agree nor disagree). We then **sum construct item responses into a *construct score* ranging from -9 to +9, with the meanings of signs, values and 0 still holding**. We import the data into IBM SPSS v29 [207] for statistical analysis. We **compare mean construct scores via One-Way ANOVAs** followed by **Tukey Honest Significant Difference (HSD) tests to confirm group differences at significance $p < 0.05$**. We further investigate the groups' scores via the **median and modes of each Likert-item to get a general sense of how the groups respond to different items, and how those contribute to the overall score**.

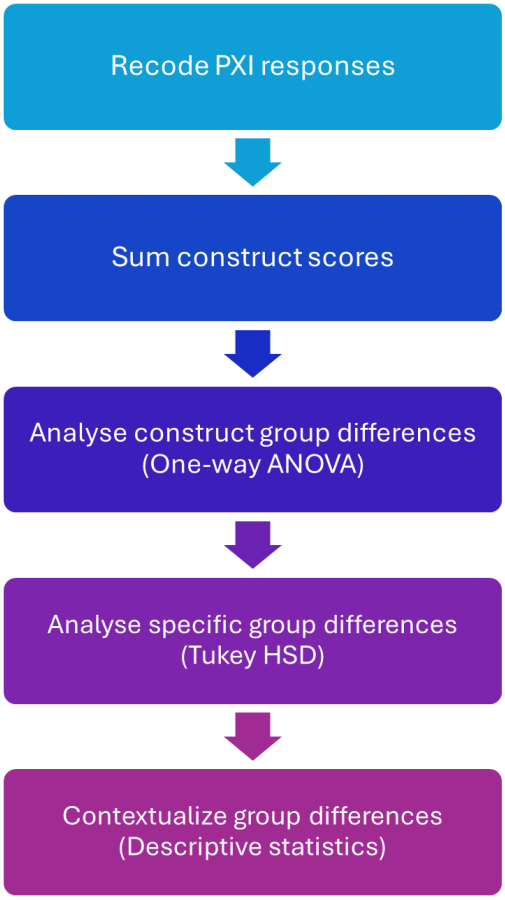


Figure 18.2: PXI Analysis Method

18.4.2 Expected Trends

We expect Easy responses will be low in Meaning, Immersion, and Challenge as the game is too easy (bordering on boring), but the Master scores will likely be high. Comparatively, Hard scores should be low across the board (as the game is virtually impossible), and Control should be high (as the game meets player skills). We derive these expectations from our understanding of PX (recall the larger survey in Part I), and summarize them in Table 18.3.

	Mean.	Mast.	Imm.	Prog.	AV	Cha.	Con.	Goals
Easy	↓	↑	↓	–	–	↓	–	–
Control	↑	↑	↑	↑	↑	↑	↑	↑
Hard	↓	↓	↓	–	–	↓	–	–

Table 18.3: Expected trends in responses.↓= low values, ↑= high values.

18.4.3 PXI Results

We focus on Meaningful, Mastery, Immersion and Challenge as we expect these to change depending on group. As there is a lot of data, we summarize important findings and what they mean in Table 18.4. We present the full results in App. H.4, and point to the appropriate tables in the body of this results section.

The results are counter-intuitive to our expectations: Easy's mean responses are the highest across constructs (Table H.13), Control and Hard never statistically differ from each other (Table H.15), and while Hard does have some negative means (Meaningful, Mastery) they are fairly close to neutral (see Table H.16 for details). We find **significant group differences for Meaningful *($F_{2,74} = 6.26$, $p < 0.005$)* and Mastery *($F_{2,74} = 6.93$, $p < 0.005$)***. Meaningful and Mastery means differ between groups (Hard: negative, Control: neutral, Easy: positive). In Meaningful, Easy differs from both Control (Diff: 2.93, $p < 0.05$) and Hard (Diff: 4.05, $p < 0.005$); but Control and Hard are not statistically different. In Mastery, Easy and Hard differ (Diff: 5.10, $p < 0.001$), but no other pairing is significantly different. **Challenge was close to being significant between groups *($F_{2,74} = 3.03$, $p = 0.054$)***. Challenge's group means were all positive; looking at individual pairings only Easy and Hard significantly differ (Diff: 3.07, $p < 0.05$). **Immersion is not significant, and its means are positive across groups**.

18.4.4 Discussing PXI Results

Meta-Explanations: Groups are randomized, so within each group there may be participants who felt the over- and under-loading to greater degrees. Their responses may pull the group averages in ways that made constructs seem not significantly different. It is also possible that a player's identities and performance expectations could skew their responses based on item wording.

Meaningful: It is unclear why Easy feels somewhat meaningful (Mean: 3.13), while the other two feel neutral (Control Mean: 0.19, Hard Mean: -0.92). It could be that the **individual items were ambiguous in defining "meaningful" (Mean1), "relevant" (Mean2), and "valuable" (Mean3)**. Participants may address this ambiguity by thinking about their scores between the baseline and experiment as an evaluation of meaning. **Perceived score changes could inspire positive (if rank increases) or negative (if rank decreases) feelings which participants interpret as "meaningful"**. This could explain why Control is fairly neutral, but not why Hard is also neutral. Considering modes, most Control participants did not find the experience meaningful (-6 - disagree), while Hard

	Results Summary	Interpretation
Meaningful	• **ANOVA**: 6.26***	• Easy perceives more meaning
	• **Tukey**: Easy/Control = 2.93*	• Control and Hard are neutral
	• **Tukey**: Easy/Hard: 4.05***	• Individual responses wildly vary
	• **Construct Desc**.: Easy means are positive and larger than other groups	
	• **Item Desc**: Values support other findings	
Mastery	• **ANOVA**: 6.93***	• Easy perceives significant mastery.
	• **Tukey**: Easy/Hard = 5.10****	
	• **Construct Desc**.: Easy's mean $\pm\sigma$ completely overlaps Control mean $\pm\sigma$	• Control and Hard perceive neutral to somewhat positive mastery.
	• **Item Desc**: Easy generally strongly agrees; Control and Hard are more neutral	
	• **Item Desc**: Control and Hard have noticeably lower responses for Mast3 than other Mastery items	
Immersion	• **ANOVA**: n.s	• Groups are similarly immersed.
	• **Tukey**: n.s	• All central tendencies trend positive
	• **Construct Desc.**: Values are extremely close and have large σ	
Challenge	• **ANOVA**: n.s (p = 0.054)	• Easy perceives the challenge as just right for them.
	• **Tukey**: Easy/Hard = 3.07*	
	• **Construct Desc.**: Trends positive	• Control and Hard perceive neutral
	• **Construct Desc.**: Central tendencies are different from each other within group	
	• **Item Desc.**: Cha3 is significantly different between Easy and Hard	
Significance: not (n.s), $p < 0.05(^{*})$, $p < 0.01(^{**})$, $p < 0.005(^{***})$,$p < 0.001(^{****})$		

Table 18.4: Important results from PXI analysis and what they mean.

participants did (+2 - somewhat agree). This further complicates the hypothesis that responses are due to the change in score/rank. **Alternatively, it could be that participant interpretations are just diverse, and so Meaningful just captures their overall attitude towards the experience**.

Mastery: Given that Control and Hard are not significantly different, it is counter-intuitive that Easy and Hard are significantly different, but not Easy and Control[5]. Looking at the means, we see our expected experiences: Easy feels Mastery, Control is neutral, and Hard does not feel Mastery. However, the **Mastery medians (Easy: 6, Control: 2, Hard: 0) are quite different from the means (Easy: 3.54, Control: 0.50, Hard: -1.56)**. The medians and modes suggest Easy feels Mastery strongly, Control feels some Mastery, and Hard is neutral. Looking at the individual items, it seems Control and Hard are greatly influenced by Mast3 (Hard - median: -2, mode: -3; Control - median: -1, mode: -1). Mast1 and Mast2 ask participants if they felt "capable" and "good" at the game; Mast3 asks if they felt "a sense of mastery". **It could be that the more positive responses to Mast1 and Mast2 are due to players not reflecting on their particular experience, but their overall feeling about how good they *think* they are at this *type* of game.**

Immersion. The group positive and not significantly different means (Easy: 4.42, Control: 3.35, Hard: 2.56) indicate all groups felt some level of immersion. Medians for individual items were similar between groups and across questions. It is possible that responses are so similar because the definition of "immersed" is ambiguous. **Participants who perceive immersion as being focused on the game may be biased because the mini-games are short and mechanically intense, thus emulating conditions of immersion in larger settings**.

Challenge. Easy (mean: 3.63, median: 5, mode: 6) seem to feel the game is appropriately tuned to their level. Control group is more split (mean: 2.62, median: 4, mode: 0); they seem to agree that the game was appropriately tuned in Cha 1 and Cha2 (Mode: 2, Median: 1), but were less sure when asked to evaluate if the tuning was to their level (Mode: 1, Median: 1). **Hard (mean: 0.56, median: 0, mode: 5) suggests that they feel neutral, but the mode suggests many believe the game is appropriately tuned for their levels.** This is odd considering it was tuned to be impossible for the players to beat, and all but one[6] member of the Hard group received either a "D" or an "F" rank in their experimental conditions across the button mashing games (an average of 2 ranks lower than their baseline). Looking at individual items, Hard responses to Cha3 (median: 0, mode: -2) deviate from the other groups. Cha1 and Cha2 ask more generally about the game, while Cha3 directly asks the player if the challenge was the right level for them. It is possible that the generic wording of Cha1 and Cha2 cause the participants to compare this game against other games they have played, resulting in a more positive score since these are "simple".

[5]We understand this is mathematically because Easy's mean and standard deviations completely overlap Control's mean and standard deviation. Here we focus on why players may respond that way.

[6]P21 received an outlying "C" rank in AIBM, but a "D" in SIBM and "F" in MIBM.

18.4.5 Summarizing PXI

From the PXI results we see:

- Players in the Easy condition have a more positive experience with the game, while believing it is appropriately tuned to their abilities.
- Players in the Control and Hard conditions report neutral feelings towards the game tuning regardless of actual performance.

These findings run counter to Flow theory [103] and other PX constructs (e.g. immersion [127]) that suggest players are disengaged from "boring" or "trivial" gameplay. As well, the fact that Control and Hard report neutral feelings regardless of degree of failure indicates that something is going on in the self-reporting.

18.5 Qualitative Strand: Textual Responses

18.5.1 Thematic Analysis Process

We adopt a "Big Q" perspective, relying on our theoretical and conceptual frameworks (Sec. 18.1) and positionality (Sec. 18.2.1) to reflexively guide our interpretations. Fig. 18.3 is an overview of our thematic analysis process. These steps are non-linear; we move back and forth, and combine stages during the holistic analysis process. Our understanding of each phase are shaped by work from Braun and Clarke [59] and Terry et al. [469] and Sasha's previous grounded theory work that used thematic analysis to study changes to gamer identities and labels over COVID-19 [60]. We detail each step below, and provide examples where appropriate.

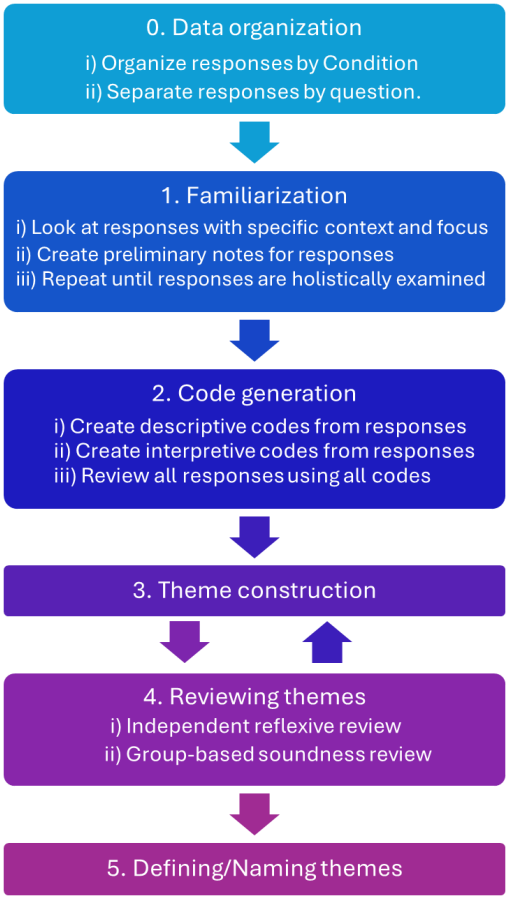


Figure 18.3: Thematic analysis process overview.

0. Data Organization. We group responses by their experimental condition (participant distribution in Table H.11). Groups are approximately the same size, average age, gender distribution (see Table 17.1 for demo details), gamer-to-non-gamer ratio (Easy 18:6, Control 20:6, Hard 16:9), and frequently reported gaming history (Easy/Control/Hard: 10+ years). We then separate the data by challenge, so we can analyse all of the SIBM Easy condition responses before moving on to the SIBM Hard responses, etc. This allows us to note codes, patterns, and observations unique to a game-condition pair, enabling more meaningful latent understandings. We treat the "final thoughts" separately and review/code it after the challenge-based responses.

1. Familiarization. We make three complete passes through our data, each with a different response context (anonymous, gamer demographic, performance). Each pass requires multiple readings to make sure that we capture as much information as possible, especially as ideas from later responses may change opinions and contexts of earlier ones.

1.1 Anonymous pass: We look at the responses with no identifying participant data. We focus on semantic meanings. We draw some initial interpretations and highlight keywords that could be relevant to coding.

1.2 Gamer demographics pass: We look at the responses alongside participant demographics (age, reported gender), reported gamer identity, and gaming history (play history, frequency, session lengths, common games, and frequently used platforms). We focus on latent and contextual meanings, prioritizing understanding that reflect participant information and concepts from our theoretical framework. We update our notes for all responses and highlight potential codes.

1.3 Performance pass: We add participant performance data (raw scores, ranks, rank differences) and our observational notes to the responses. This contextualizes responses in emotional experiences and immediate gaming contexts. We update our notes to highlight how this information enhanced or changed our understanding of the data and trends.

2. Code generation. We mix deductive (descriptively reflecting what was being said) and inductive (interpreting what was being implied) coding. We start with descriptive codes and then review the responses again with an interpretive lens, potentially resulting in new inductive codes. Whenever we generate new codes, we revisit previously coded data to see whether the new codes apply. Through this process, the same phrases or segments of data may have multiple codes applied; this is a regular occurrence in thematic analysis [59, 469] and just reflects the richness of qualitative data. We list out the complete set of codes in App. H.5. We walk through this process with P69's responses to what he felt was the hardest part of the fire-starting game (SIBM):

> "*Difficulty was not an overall factor, but ensuring position of my hands to rapidly click the right arrow key at length without succumbing to fatigue*"
>
> (P69, SIBM, Easy)
> *Man, 31, Gamer for 10+ years, SIBM Rank increased by 1*

2.1 Descriptive coding: We start by looking just at the semantic meaning of the response noting down what is being signaled through the phrases as codes. For example, when P69 says "difficulty was not an overall factor" they are signaling *the game was easy for them.* We give our codes a long descriptive form (e.g. ***Game was easy for me***), and a shorthand (e.g. ***Easy***) for quicker notation. We summarize the descriptive codes for this quote in Table. 18.5.

Code	Shorthand	Part of the response
Game was easy for me	Easy	"Difficulty was not an overall factor..."
Approaching the game with a strategy	Strategy	"...ensuring position of my hands..."
Understanding input	Input	"...rapidly click the right arrow key..."
Understanding of mechanics	Mechanic	"...rapidly click the right arrow key..."
Fatigue	Fatigue	"...without succumbing to fatigue"

Table 18.5: Descriptive codes for P69's quote for SIBM Difficulty.

2.2 Inductive coding: We then go back and look at the response for latent meanings (codes in Table. 18.6). For example, "Difficulty was not an overall factor, but ensuring position of my hands..." implies P69 sees their hand position (code ***Strategy***) as having a non-trivial impact on their performance. P69's linking of a particular strategy and increased performance signals they have *meta-knowledge* ("the set of skills, strategies, and attitudes that support optimal play" [60]) for these challenges, likely developed over their gaming history (10+ years, many different genres). While they performed well, this phrase shows they are ***justifying their performance through meta-game knowledge***. This aligns with what we understand of game literacy and gamer identities.

Code	Code shorthand	Part of the response
Conceptual model of gameplay	Conceptualization	"...ensuring position of my hands to rapidly click the right arrow key..."
Justifying their performance through meta-game knowledge	Justifying performance	"Difficulty was not an overall factor, but ensuring position of my hands..."

Table 18.6: Inductive codes for P69's quote for SIBM Difficulty.

3. Theme construction. We cluster codes based on similarities in their use, and the phenomena they describe into candidate *themes*[7]. This entire subjective process relies heavily on our understanding of the data (from familiarization), our theoretical and conceptual frameworks, and our positionality. We give more weight to the inductive codes as they reflect latent meanings and provide more rich data. While "good quality themes should be distinctive, with little 'bleeding' of codes" [469, p. 28], it is possible for themes to overlap their concepts and codes in ways that suggest a unifying framework [59].

4. Reviewing themes. We break this process up into two steps: an independent reflexive review, and a group-based review.

4.1 Reflexive review: We compare candidate themes to see if they have any relationships, or if they are further collapsible into a more comprehensive theme. We double check the themes to see if they meaningfully capture aspects of data and its tone, and relate to the

[7]Underlying concepts and patterns in data which relate to our research question.

research question. If we are unsatisfied we return to theme construction. Once satisfied we move to the group review.

4.2 Group review: We present our candidate themes to other members of the Games Scalability Lab (G-ScalE) alongside coded excerpts. This serves as a soundness check of our themes and reasoning. We prime the group with an overview of our theoretical and conceptual framework so they can see whether our interpretations of the coded excerpts seem logical, and therefore fit the overall theme. If at any point the themes do not seem reasonable, we return to the theme construction stage and try a new organization.

5. Defining and naming themes. We finalize theme names, their supporting examples, and the story of their interpretation and meaning. We use excerpts from player responses as narrative evidence for the themes. We attribute excerpts to the participant via their ID, and list the game being discussed, and the condition. We also outline the relationships between themes and draw conclusions from across the whole analysis.

18.5.2 Results: Factors Influencing PX of Gameplay

Two major themes emerge from our data: the player's *conceptualization of the challenge*, and their *conceptualization of themselves*. Conceptual models[8] reflect a person's understanding of an object, experience, situation and/or person based on their previous experiences and knowledge [350]. A player's conceptual models of a game impact their in-game performance, feelings of presence, and frustration (e.g. 289, 398, 524). Their conceptual model of being a gamer influences the types of experiences they seek out (e.g. 30, 534), how they approach games (e.g. 202), and how they conceptualize failure (e.g. 11, 146). The interaction of these two conceptual models (challenge and gamer-self) form the basis of the perceived difficulty and PX of gameplay.

Theme 1: Conceptualization of the Challenge.

The conceptual model of a game covers a player's understanding of the formal game elements (e.g. goals, mechanics, rules), appropriate strategies, and general opinions on the gameplay. When encountering any gameplay, players fallback on their conceptual models to determine how they should approach the challenge and what they should expect of the game.

We find the players conceptualization of the gameplay seems to address two underlying questions: (1) *is the challenge simple or complex*, and (2) *is the challenge skill-based or strategy-based*?

Simplicity vs. Complexity: Players broadly conceive of a challenge as either *simple* or *complex*. Simple challenges are considered "easy", a perception seemingly constructed through comparing the game's formal elements to other games. This recurs across games and conditions when players say things like "*the only hard part was spamming the keyboard*" (P13, SIBM, Hard). Complex challenges require players "*to think about what [they were] doing*" (P30, MIBM, Hard). Complexity implies difficulty, and is constructed from the ways

[8] Also known as mental models.

the game's formal elements combine with the potential approaches. **Player's view of a challenge as simple or complex seems to be entirely based on gaming literacy, not in-game performance.** Consider these quotes about the difficulty of AIBM from two longtime gamers:

> *"this game wasn't really that hard, I thought it was easy as you had to press the keys one after the other"*
>
> (P43, AIBM, Easy)

> *"I think it was knowing if I should use my right hand or using both hand fingers - I like challenges so I wanted to know which way I could do it so it was challenging - maybe using just my right hand but I like to win so I liked to focus during it"*
>
> (P59, AIBM, Easy)

P43, a team-based shooter focused computer player, constructs the game as "easy" based on its mechanics and goals held in comparison to the more cognitively demanding gameplay he is used to. He sees no further depth to the gameplay or ways to engage with it beyond press the buttons fast. P59, an action-adventure focused console player, constructs the game as "complex" through strategizing whether one or two-hands would be optimal. Her response implies a more sophisticated literacy for button mashing, which could reasonably have been developed based on her age and gaming context.

Skill-based vs. Strategy-based: Players consider challenges as either *skill* or *strategy* based. Skill-based games focus on player proficiency and limitations, with discussions about how "*I was getting tired and felt like I was losing my rhythm*" (P47, AIBM, Hard). In comparison, strategy-based games are about finding "*what type of button mashing would work best*" (P22, SIBM, Control). We can consider the differences between views from the following quotes:

> *"Pressing the button fast enough (I felt like I was lacking the only skill needed to do well in the game)."*
>
> (P14, SIBM, Control)
> *Skill-based*

> *"Mashing at the same time, using one hand was more consistent but would obviously be slower than two hands"*
>
> (P47, MIBM, Hard)
> *Strategy-based*

While both views agree that competency is needed to perform well, skill-based considers it the only determinant (e.g. P14), while strategy-based believes that optimal technique overcomes competency deficits (e.g. P47). **Players often consider simple as skill-based, and complex as strategy-based.** This feeds into their underlying expectations about how much effort they should put into the game and how much experimentation is necessary. However, this is not a hard and fast rule as we see with P47 who seems to view the challenge as simple with a straight-forward approach, even though she clearly recognizes it as strategy-based.

Theme 2: Conceptualization of Themselves.

While gamers are not a monolith, they all exist inside gaming culture and absorb its values. Gaming culture sees itself as a meritocracy [60, 220, 365] which values "skill above everything else" [220, p. 79]. Gamers embody this value, and exhibit a sort of *gamer exceptionalism* (e.g. 41, 60, 94), reinforced by research about how gamers have superior abilities than non-gamers (e.g. 122, 169, 453), and are more persistent in the face of failure (e.g. 506, c.f 12).

When faced with the possibility of failure, this strong identity and sense of exceptionalism primes gamers for identity threat soothed by strong self-affirmation and self-regulation responses [37]. We see these responses emerge through: (1) asking *am I this type of gamer?*, and (2) *living up to gamer expectations.*

Am I this type of gamer? Individuals see themselves as a "type" of gamer; their "type" acts as a lens to view their performance, and set their skill expectations. Colloquially "skills" refers to the cognitive and motor abilities needed to excel at the mechanics of a game. So it makes sense that a strategy (i.e. cognitively-focused) gamer like P71 would **self-affirm this type identity in the face of perceived failure** by saying they are *"[n]ot particularly experienced with mashing games"* (P71, General Feedback, Hard). **Players using this defense may mischaracterize the gameplay to fit genres they are familiar with and are outside their type.** For example, P58 clarified they *"...don't usually do well with rhythm games"* (MIBM, Hard). While none of our games are rhythm games in their design or mechanics, P58 may perceive them that way due to her 10+ years of gaming history which focuses on games that do not involve button-mashing-as-a-challenge. This exposure to gaming culture could lead to this particular type of defense despite of not having a strong "gamer" identity.

Living up to gamer expectations:

> *"I felt that I should have spent* ***more time figuring out what techniques worked best****, but I* ***wasn't expecting to need to optimize it*** *that much because* ***button mashing games are usually pretty easy and don't have high thresholds for success****. I enjoyed the game but did not enjoy the taste of failure. Generally* ***I don't feel motivated to button mash as hard as possible unless it's in a competitive setting****, or a reward I genuinely want."*
>
> (P58, General, Hard)
> Bolded for emphasis.

In failure, we see the underlying expectations and reactions that gamers have about the challenges (Quote 18.5.2). Many gamers assume inherent competency regardless of knowing their abilities were being tested. Faced with failure, their responses rely on common gaming rhetoric to distance themselves from their performance and self-regulate their identity. We see defensive biases like describing ways they were knowingly sub-optimal (e.g. not figuring out the best technique, not going as hard as possible), and attributing blame to the game (e.g. not feeling motivated because it was not competitive/did not have a reward they wanted, not having the right technology, not having the right type of feedback). In most of these

identity threat responses there is palpable frustration as gamers tried *"not getting angry at a bad rating"* (P74, All Games, Hard).

18.5.3 Summarizing Themes

Theme	Sub-themes
Conceptualization of the Challenge	Simplicity vs. Complexity
	Skill-based vs. Strategy-based
Conceptualization of Themselves	Am I this type of gamer?
	Living up to gamer expectations

Table 18.7: Themes and subthemes from Thematic Analysis.

Table 18.7 summarizes our themes: two conceptual models (challenge and gamer-self), and their implicit questions, that provide insight into a player's expectations about and experience of gameplay. These conceptual models are not mutually exclusive, as game literacy (and by extension the conceptualization of the challenge) is informed by the player's relationship to the gamer culture and identity. **Misalignment of their expectations create frustration**. Gamers self-regulate with defensive behaviours like attributing performance issues to the game, or self-affirm by distancing themselves from the type of game, or explaining how the circumstance is less than ideal (e.g. not having the right technology). **These behaviours save face because it allows them to be "good gamers" even though they did not succeed**. Non-gamers seem to self-attribute failure, likely because they do not expect themselves to succeed. **Regardless of actual skill-difficulty tuning, if a player believes they should have "succeeded" and they do not, their view of the experience will be negative**.

18.6 Integrating PXI and Thematic Results

We integrate our analyses, using our themes to contextualize why PXI findings run counter to our expectations and conventional knowledge.

Why does Easy have high PXI scores? Overall players' perceptions of the experience seem directly related to their success in the game. Players in the Easy condition were always successful, and did not question this behaviour. Their conceptual models aligned, so they expected high performances. The high levels of Meaningfulness reported by the Easy condition seem to reflect this alignment, implying that meaningfulness may be a construct measuring expectations, and so is heavily influenced by positive experiences (i.e. winning). This aligns with Bowey, Birk, and Mandryk [53], which finds that feelings of success increase the player's view of their own competency, autonomy, presence, enjoyment, and positive affect.

Why do Control and Hard not differ? The Control and Hard groups expectations are not being met. Previous work found that gamers seem to have a more nuanced interpretation of "what counts" as failure focused more on their self-expectations [11, 146]. This could explain why PXI measures for these groups were not significantly different since both groups view success as achieving an S-rank. Hence a less than perfect score (Control group) and failing score (Hard group) are equally bad. The lack of difference in intensity, while surprising, could be a reflection of player's responding to stereotype threat. This would explain why the Control and Hard group's were unsure about the amount of feedback since that is a common way poor performance is attributed to the game.

Why are Control and Hard responses neutral? The PXI responses seem to reflect the player's conceptual models rather than the specifics of the game they just played — hence why we see players who failed rating the game as "not too easy/too hard" (Cha1). This would also explain why we see players from the Hard group consider themselves to have some amount of Mastery at the game (see Mast1 and Mast2), in the face of failure. They believe their conceptual model is correct, and so may be responding to these questions as if they were asking "are you good at button mashing" instead of directly regarding these games.

Overall. Our findings lend some support to the idea of the "Impression Manager" player type [183]. These are players who are motivated by the social capital of being "good" at a game but are so risk-averse that they only want to play at difficulty levels that guarantee their success. Impression managers therefore want to engage with the rhetoric that games are a skill-based meritocracy since it gives them prestige, but they want to always be perceived as having significant merit. This kind of motivation could be exacerbated by the social hierarchies and rhetorics of competitive gaming culture [60], which are now the dominant gaming culture. This reinvigorates the case for dynamic difficulty adjustment in games, especially based around their competency profiles, as it could be particularly effective for this type of gamer.

18.7 Conclusion for this study

Our mixed-methods study results are short:

- Player's perceived experience is influenced by their expectations, not just their skill-difficulty balance.
- Player expectations are the result of their *conceptualization of the challenge* and their *conceptualization of themselves as a gamer*.
- Aligned expectations skew positive experiences and belief in game balance.
- Misaligned expectations skew negative experiences and can induce gamer identity threat.
- Gamer self-affirmation could impact the measurement of experience constructs for negative experiences.

This aligns with existing game studies literature on gamer identities and player types, even though it disagrees with PX construct literature (i.e. Flow and immersion). We can interpret

these results as showing that mechanical experiences (even the limited scope that we are looking at) are a single factor in larger PX, with the results suggesting that in this case the socio-cultural experiences may have a stronger influence on the holistic PX.

What this means for our thesis. RQ3 asks how designers can use the knowledge of a challenge's mechanical experience in their game design. Jutsus as knowledge capture tools visualize the mechanical achievability of a specific challenge for a particular player making it easy for designers to tune challenges for desired PX. Expressing mechanical experiences through mechanical achievability, we rely on understandings of Flow and challenge-based immersion to dictate what challenge tunings would result in "good" PX. We originally envision the competency profiles as representing the centre of an ability range that could possibly complete a challenge and so feel positive PX. Seeing how players do not actually want perfectly challenging experiences, means that we should interpret challenge competency profiles as ability floors necessary for a positive PX. This must be kept in mind as we design jutsus and explain how to interpret them.

Limitations in Study. We discuss generic limitations in Ch. 15.5. A unique limitation of this study is that we took the participant's PXI measurements after having played all the games so their responses *may* reflect their experience with *all* of the button mashing games they played. Having participants fill out PXI forms after each challenge would significantly increase their experimental time and could make their results less reliable since they would be spending 10 seconds on the game and about 6 minutes for the survey (the average amount of time it took to complete). By having the survey at the end, participants had more time in their experimental condition to reflect on their feelings and so their responses give us a sense of how the tuning relates to the experience.

Future Work. We should explore this idea through dynamically adjusted experimental conditions to see whether there are stronger or different results when participants are grouped by ability level. Studies like this would benefit as well from a more in-depth qualitative process that targets how integrated a player's gamer identity must be before these effects appear. Future studies should also consider recording participant player types to see whether there are deviations along those motivation lines.

Take home points

From this chapter we learned the following meta-lessons:

- Self-report measures for experience are susceptible to bad data when a negative experience triggers the respondent's identity threat.
- Player's experiences are the result of their expectations of the game and themselves (positive = aligned expectations; negative = misaligned expectations).
- Players want to succeed more than they want to play a balanced challenge.
- Our jutsus should reflect this by highlighting competency profiles as ability floors.

Chapter 19

Closing Remarks: The Challenge Model

We close out Part III with some promising preliminary answers for our research questions. RQ1 is addressed by the challenge models for our Button Mashing family of challenges (Tables 19.1, 19.2, 19.3 on the next three pages) based on their hypothesized (Ch. 14) and validated (Ch. 16) competency profiles. RQ2 is addressed by our competency profiles at different loadings (Ch. 17), which show us that performance only uniformly scales with limiting ability when there are no "important but not limiting" abilities at play. We start to address RQ3 through Ch. 18, and end up reminding ourselves that the perceived experience is more than just the mechanical experience. Through all of this we have a more robust understanding of the mechanical experiences of button mashing challenges through the lens of limiting abilities. This is an important stepping stone to improving our overall understanding of mechanical experiences and the effectiveness of our challenge model.

19.1 Improving our work

Future work in understanding the relationship between human abilities and atomic challenges in a quantitative way should start by improving their measurement methods for abilities. We foresee this becoming a more limiting factor in validating challenge competency profiles, especially for cognitive-focused challenges where abilities overlap significantly. Most challenges in commercial games require more significant combinations of motor and cognitive abilities than we see here, and some flip the script to rely more heavily on cognitive abilities entirely. Having more precise and reliable measurement tools will allow for more specific understanding of our more complex challenges.

With an improved measurement method, the next conceptual step in this type of research is to understand how to account for or incorporate elements of a player's gaming literacy and experience into these kinds of models. We see that even in a motor-dominant challenge family like Button Mashing Games, players interpret the games differently and employ different strategies which may be significantly impacting their performance. This makes the relationship between the challenge and abilities somewhat suspect as a person with lower ranked abilities may be able to perform well due to strategy differences. While this is minimized in

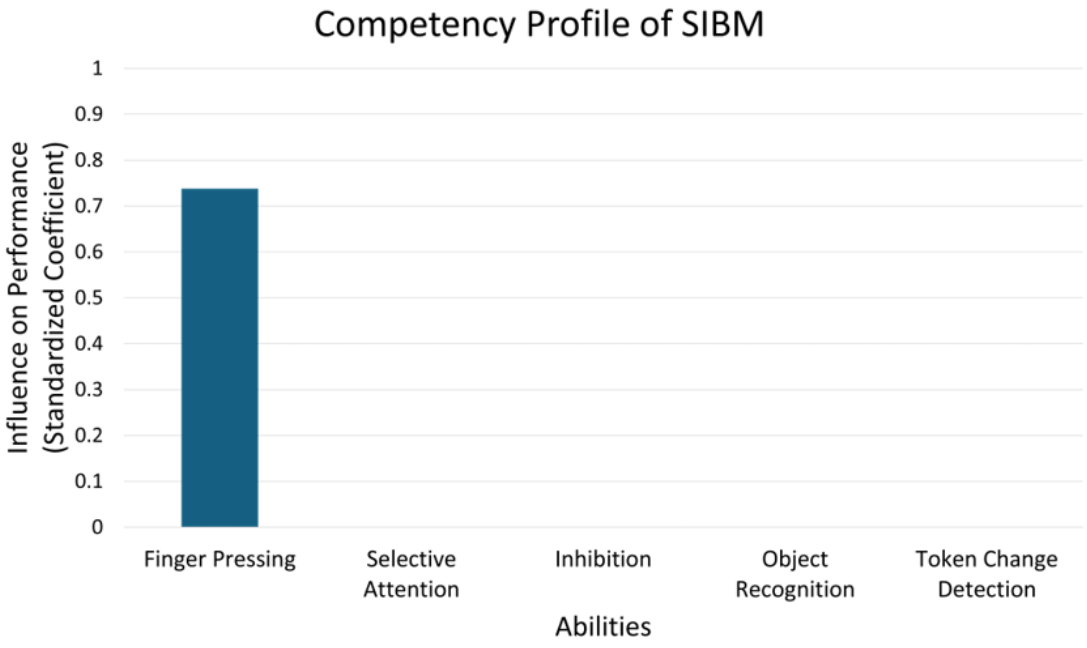
Name	Single Input Button Mashing
Challenge Information	
Definition	Repeatedly pressing a specific single button or key on the controller as fast as possible within a given time limit.
Mechanics	
Button pressing, short time limits	
Variable components	• Target button • Time limit • What "counts" as a press
Context	
Interaction	Pressing buttons
Controller Type	Standard Controller
Number of Players	Single player
Type of Play	Competitive (solo)
Examples	
• Dragon's Breath, South Park: The Stick of Truth • Torture Attacks, Bayonetta and Bayonetta 2 • Boss Knockouts, Donkey Kong Country: Tropical Freeze	
Variants	
• Controller Type: Handheld console, handheld motion controller (horizontal), keyboard • Number of Players: Multiplayer • Type of Play: Cooperative-Competitive (team based), Cooperative	
Intrinsic Competency Profile	
	
Source of Difficulty	Finger pressing

Table 19.1: Challenge Model for Single Input Button Mashing Game.

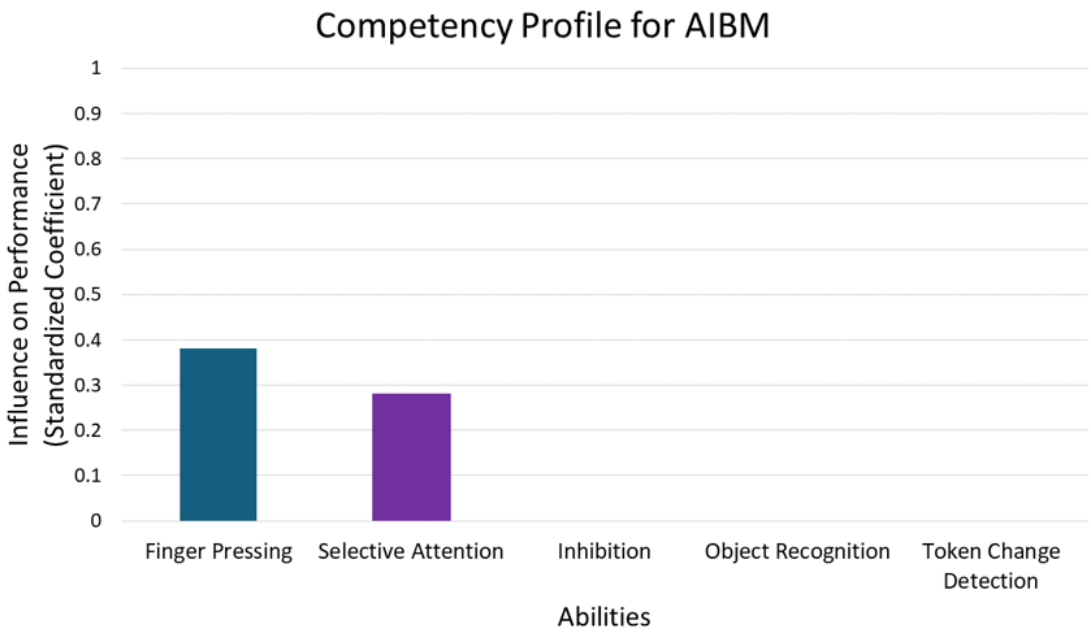
Name	Alternating Input Button Mashing
Challenge Information	
Definition	Repeatedly and rapidly press two specific buttons in sequence.
Mechanics	
Button pressing, short time limits, pattern recognition	
Variable components	• Sequence length • Target buttons • What "counts" as a press • Time limit
Context	
Interaction	Pressing buttons
Controller Type	Standard Controller
Number of Players	Single Player
Type of Play	Competitive (solo)
Examples	
• Take a Breather, Mario Party 4 • Balloon Burst, Mario Party • Psychic Safari, Mario Party 2	
Variants	
• Controller Type: Handheld console, handheld motion controller (vertical), handheld motion controller (horizontal), keyboard • Number of Players: Single player • Type of Play: Cooperative-Competitive (team based), Cooperative	
Intrinsic Competency Profile	
	
Source of Difficulty	Finger pressing

Table 19.2: Challenge Model for Alternating Input Button Mashing Game.

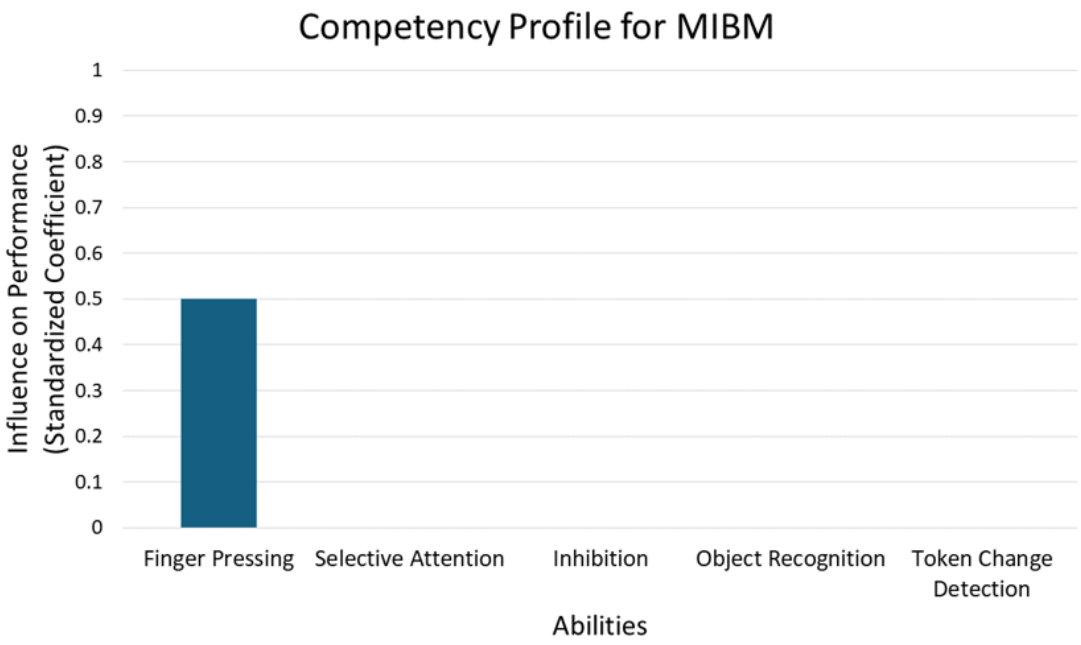
Name	Multiple Input Button Mashing
Challenge Information	
Definition	Pushing multiple buttons simultaneously, repeatedly, and rapidly.
Mechanics	
Button pressing, short time limits	
Variable components	• Number of buttons pressed • Target buttons • What "counts" as a press • Time limit
Context	
Interaction	Pressing buttons
Controller Type	Standard Controller
Number of Players	Single player
Type of Play	Competitive (solo)
Examples	
• Mecha-Marathon, Mario Party 2 • Chin-up Champ, Wii Party	
Variants	
• Controller Type: Handheld console, handheld motion controller (vertical), handheld motion controller (horizontal), keyboard • Number of Players: Single player • Type of Play: Cooperative-Competitive (team based), Cooperative	
Intrinsic Competency Profile	
	
Source of Difficulty	Finger pressing

Table 19.3: Challenge Model for Multiple Input Button Mashing Game.

motor-dominant challenges due to biomechanical limitations of human movement, we foresee it becoming problematic in cognitive-dominant challenges.

19.2 Wrapping Up

What we learned in this part:

- Challenge performance is reasonably modeled as a linear combination of some subset of human cognitive and motor abilities (the competency profiles being direct answers to RQ1).
- Our method for creating competency profiles seems to produce reasonable approximations of the actual competency profiles.
- Our understanding of limiting abilities as main drivers of success is reasonable when considering over and underloading challenges (RQ2).
- Performance in atomic challenges is susceptible to influence from gaming literacy/experience (i.e. strategy)
- Even in pure skill-based gameplay, player's perceived experience is more related to their expectations of the game and themselves as gamers than pure mechanical experience (related to RQ3).

What we produced in this part:

- A method for creating challenge competency profiles that include cognitive abilities;
- Expanded and validated competency profiles for single-input, alternating-input button mashing games (see charts in Tables 19.1, 19.2);
- Expanded and semi-validated competency profile for multiple-input (see chart in Table 19.3) button mashing games;
- Three custom button mashing games for testing purposes (App. D);
- A method for experimentally validating competency profiles; and,
- Final challenge models for our Button Mashing family of challenges (Tables 19.1, 19.2, 19.3).

Part IV
The Jutsu Framework

We have finally arrived to the main event of this thesis: combining our player model (Part II) and challenge model (Part III) into *jutsu*.

Through Ch. 20 we elaborate on *what* a justu is (Def. 4.1), specifying *how* it is constructed and interpreted by presenting jutsu for our button mashing challenges. Using these examples, we build a case for the usefulness of jutsu in helping designers assess a challenge's mechanical achievability and potential experience. This chapter serves as our answer to RQ3, and acts as an invitation for future researchers to continue this work. We end this part, and our thesis, in Ch. 21 with a reflection on what we have learned, and short walk through new areas of investigation.

Chapter 20

Jutsu

Back in Ch. 4.3 we defined jutsus as structured representations of a specific *mechanical experience (MX)* made from combining a *challenge model* and a *player profile* (Def. 4.1). Jutsus visualize a player's MX of a challenge by highlighting mechanical achievability and sources of mechanical difficulty (i.e. design elements that affect the experience). They are intended to be used by designers to tune their challenges towards particular experiences for individual players or demographics of players.

We begin by explaining how to construct and interpret a jutsu's mechanical experience analysis graph. We then present the jutsus for our button mashing challenges, highlighting what we can learn about the challenge designs from the jutsus. We end by proposing hypothetical scenarios to illustrate ways jutsus can be used by designers.

20.1 Constructing Jutsu

Recall Ch. 4.3 explained the structure of a jutsu as three parts: the challenge description, player profile, and MX analysis. Challenge descriptions are constructed through our game reading techniques (Ch. 13), and player profiles are constructed through our ability battery (Ch. 12). We do not repeat that information, and instead focus on the construction of the mechanical experience analysis.

A Quick Aside...

Our process is based on and illustrated by our button mashing challenges. While we believe the construction steps are applicable to other challenges, they may require further refinement or tweaking.

20.1.1 Mechanical Experience Analysis

To visualize MX we need a specific player profile and challenge instance for our computations. Let us consider P22 (Fig. 20.1) and our experimental SIBM implementation (Table 20.1).

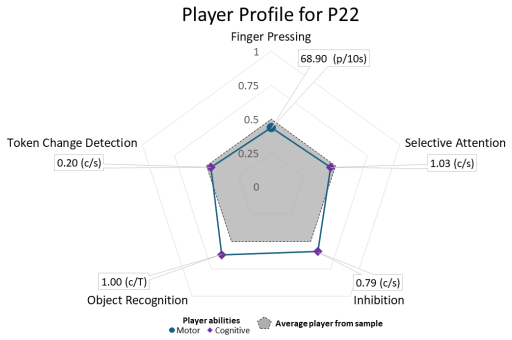


Figure 20.1: P22 Profile. Blue dots indicate motor abilities, purple dots indicate cognitive abilities. ***Legend:*** *p: Button presses, c: Correct Responses, T: total number of trials, s: Seconds*

Model:	$26.70 + 0.68 \times \text{FP}$	
Mechanic	**Variable**	**Value**
Target Button	Button	→
Time limit	Time	10 seconds
Press "counts"	Score Modifier	1
	Goal	77

Table 20.1: Fire Starter (SIBM) Information for Jutsu. ***Legend:*** *FP: Finger Pressing*

1. Gathering Challenge Information

Finding the ability minimums. We do this using gameplay details. For SIBM with score modifier of 1, our calculation is:

$$\text{Goal} = 77 = 26.70 + 0.68 \times \text{Finger Pressing}$$
$$\text{Finger Pressing} = 73.78 \text{ presses/10 seconds}$$

Finding abilities for ranks. We calculate minimums based on each rank's lower threshold. Recall for our games, meeting the goal is the lower end of B-rank, with ranks going up or down by 0.25×Goal. SIBM calculations are summarized in Table 20.2.

Rank	Score Threshold	Finger Pressing
S	115.5	130.25
A	96.25	102.01
B	77	73.78
C	57.75	45.55
D	38.5	17.31
F	0	0

Table 20.2: Minimum Finger Pressing Ability for each SIBM Rank, rounded to two decimal places.

2. Gathering Player Information

Finding error bars. Player performance is not guaranteed to be consistent between games. We represent this potential ability range through using the standard deviations (σ) of the

sample. The low end (Measure -1σ) represents a "bad day" where the player is underperforming; "good days" (Measure $+1\sigma$) similarly mean overperforming. Table 20.3 summarizes P22's ability ranges.

Ability	**Measure**	**-1σ**	**+1σ**
Finger Pressing	68.90	55.85	81.94
Selective Attention	1.03	0.79	1.26
Inhibition	0.79	0.67	0.91
Object Recognition	1	0.95	1.05
Token Change Detection	0.20	0.16	0.23

Table 20.3: P22's ranges for abilities, rounded to 2 decimal places.

3. Plotting Analysis Graph

We plot the player profile as a bullet graph using two axes (blue for motor abilities, and purple for cognitive abilities). The player's potential range is indicated via error bars. Behind the player information we display the minimum limiting ability measure for the challenge instance, along with the rank that the player would achieve at different ability scores. Fig. 20.2 is the ***mechanical analysis graph*** for our example.

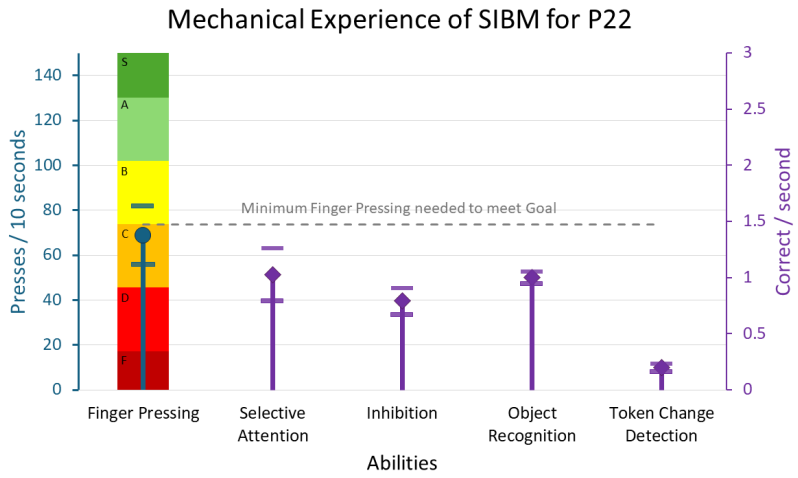


Figure 20.2: Mechanical Experience of SIBM for P22

Interpreting the Analysis Graph. We can immediately see this SIBM tuning may not be achievable for P22, as their finger pressing ability barely meets the minimum. Considering their range, we see they are more likely to fail this challenge than succeed (though success is possible). In light of what we know about perceived experiences (see Ch. 18) this tuning looks like it will frustrate P22 and will not be enjoyed.

What about multi-ability challenges?

We treat secondary abilities as thresholds that need to be met in order for the performance to rely on the limiting ability. We choose a handful of meaningful thresholds for the secondary ability and generate analysis graphs at each value. Thresholds are represented by light-gray boxes behind their associated ability. Fig. 20.3 shows P22's potential mechanical experience for AIBM at three selective attention thresholds (mean, and $\pm 3\sigma$).

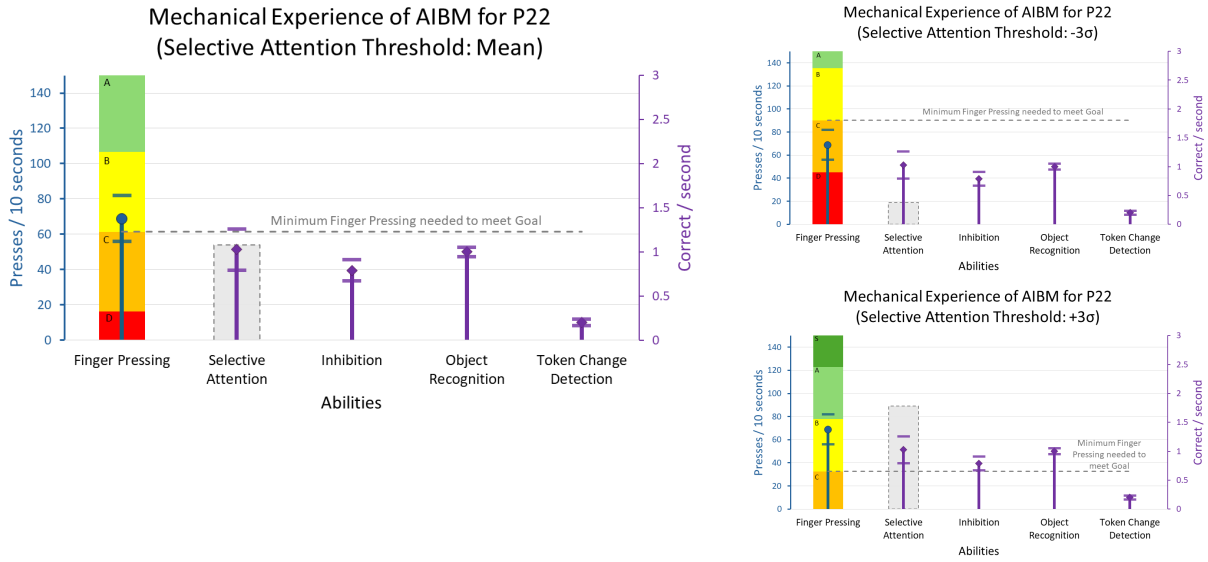


Figure 20.3: Comparing P22's AIBM Analysis graphs at Thresholds: Mean and $\pm 3\sigma$

Interpreting Multi-Ability Graphs. Ranks and minimums must be understood as being calculated *given the secondary threshold is met*. In order to use the graphs to assess mechanical achievability we need to find the spot where the player's secondary ability meets the threshold **and** where their secondary ability's lower range touches the threshold. For P22, their AIBM mechanical experience is expressed between the Mean and $-\sigma$ thresholds (Fig. 20.4). Between these graphs we see that P22 is likely to get a low B-rank on average, though they may drop significantly on "bad days".

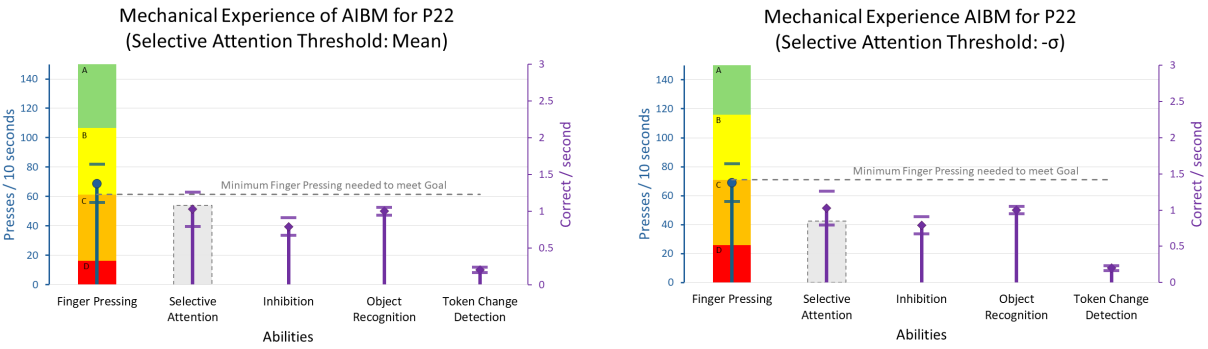


Figure 20.4: P22's Mechanical Experience of AIBM (Thresholds: mean, $-\sigma$)

20.2 Presenting Jutsus: Button Mashing Challenges

The following jutsu are constructed using our experimental games tuned to the baseline (challenge descriptions in Tables 19.1, 19.2, and 19.3), and a player homunculus representing the "average player" from our competency profile study sample (Fig. 20.5). For readability we only present the mechanical experience analysis graphs.

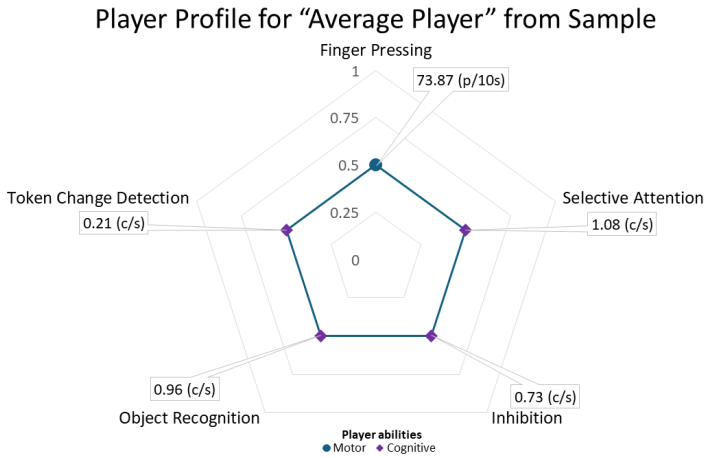


Figure 20.5: "Average Player" Homunculus: Profile representing a fake player with the mean score in each ability.

20.2.1 Single Input Button Mashing

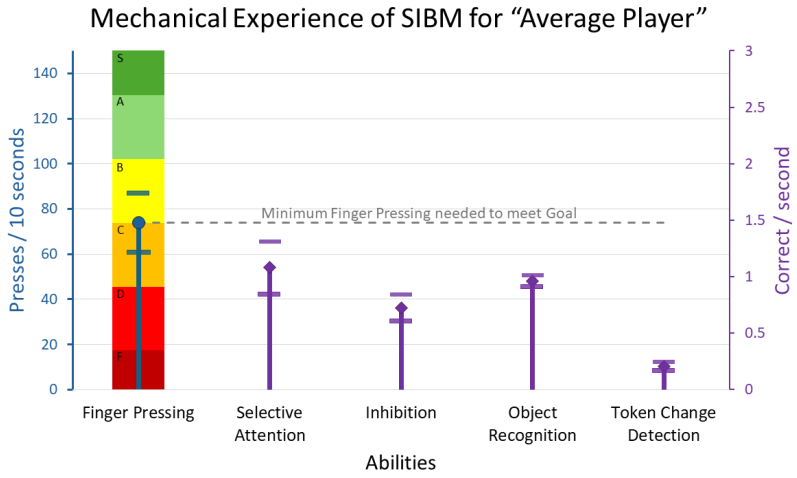


Figure 20.6: SIBM Mechanical Experience for Average Player.

From Fig. 20.6 we see the average player is reasonably challenged by the current tuning, as it becomes a 50-50 chance of success (B vs. C-rank). While this in theory is "fair" and

"balanced", we understand players will dislike this because they expect an A-rank or better. The jutsu confirms that the gameplay was mechanically reasonably tuned, and therefore response was related to perceived difficulty being too much.

20.2.2 Alternating Input Button Mashing

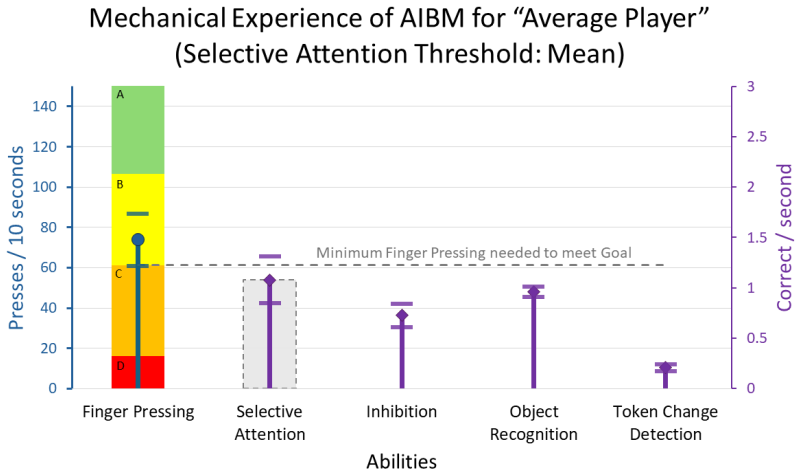
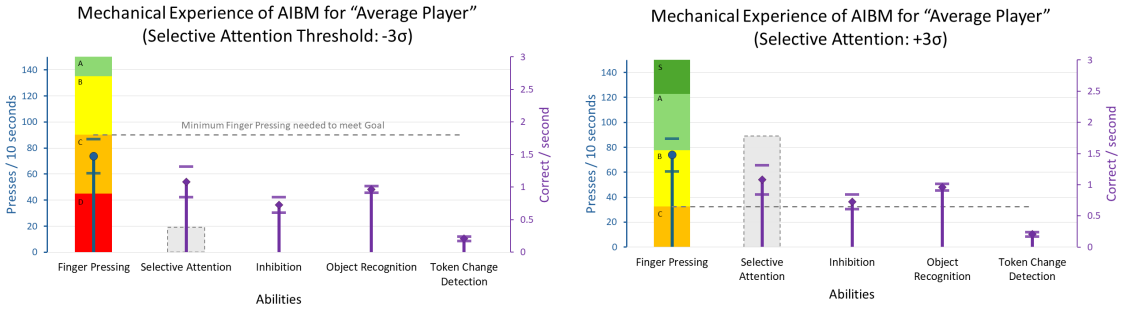


Figure 20.7: AIBM Mechanical Experience for Average Player at Threshold Mean.

From Fig. 20.7 we gather AIBM may be somewhat easy for the average player, as they will always clear the goal. This tracks with the comments left by players in Ch. 18 that it was simultaneously the "easiest" and most enjoyable of the three challenges. Looking at the extremes (Figs. 20.8) failure is possible given the trade-off between finger pressing and selective attention. However, it is impossible to get an F-rank and seems genuinely difficult for the average player to get a D-rank.



(a) (b)

Figure 20.8: AIBM Mechanical Experiences for Average Player at Thresholds $\pm 3\sigma$.

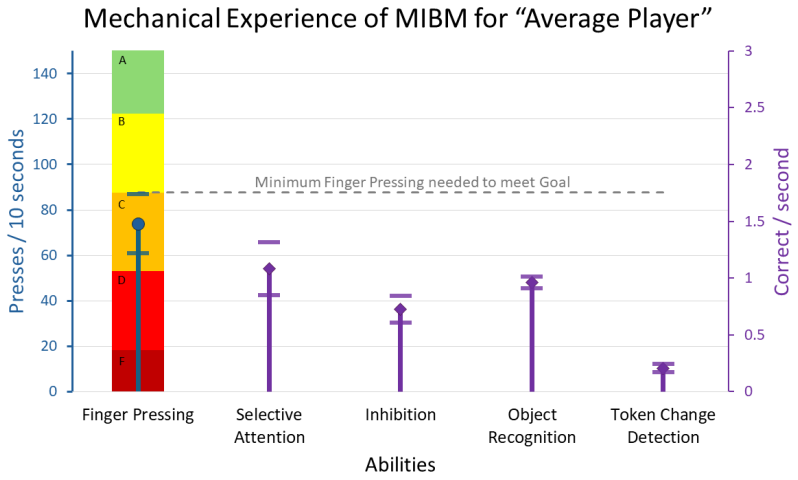
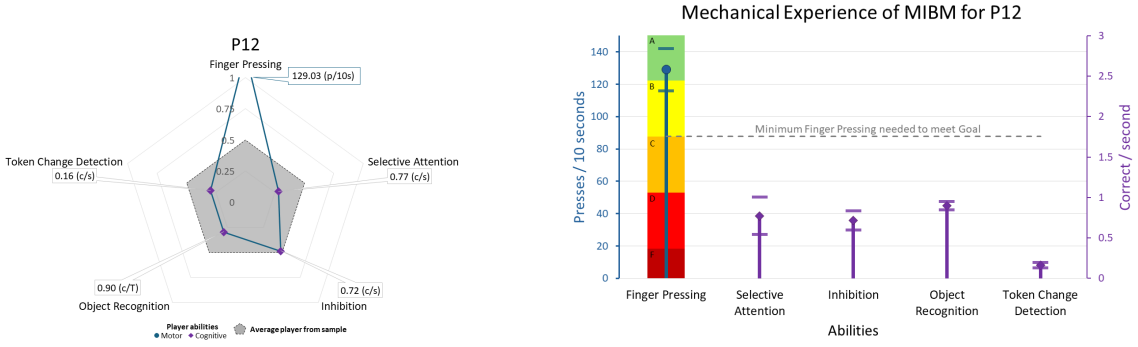


Figure 20.9: MIBM Mechanical Experience for Average Player.

20.2.3 Multiple Input Button Mashing

From Fig. 20.9 we see the game is virtually impossible for the average player, whose abilities barely meet the minimum requirements on a "good day". The only participants who can reliably clear the challenge are top-performers, with even our best Finger Presser (P12 - Fig. 20.10a) only able to barely break A-rank (Fig. 20.10b). S-ranks are not possible. This overall makes it clear MIBM's base tuning was too difficult.



(a) (b)

Figure 20.10: MIBM Mechanical Experience for P12.

20.3 Tuning Gameplay with Jutsu

Visualizing button mashing MXs through jutsus immediately highlights the mechanical achievability and its affect on the perceived experience. While this is useful, we can go a step further by using jutsu to tune the difficulty (and thereby MX) of our challenges. Using our challenge models, we can "test" different game tunings against player profiles without

having to touch the actual game. The jutsus then display what this new experience would be, allowing for interactive design tuning.

Consider SIBM, our original baseline tuning (Goal: 77, Score Modifier: 1) felt unfair to players who found they were not doing as well as they would expect. If we want to align the gameplay to our player's expectations we could modify the intrinsic difficulty of the challenge, or just the perceived experience of the challenge.

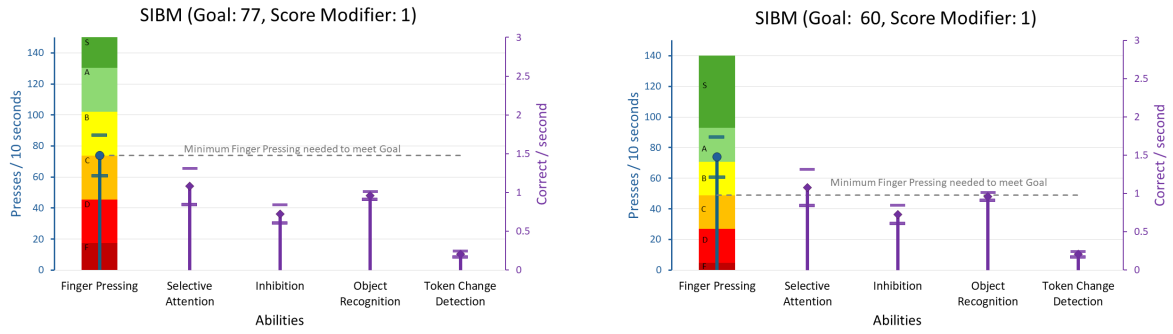
A Quick Aside...

Changing the intrinsic difficulty of a challenge will change its perceived difficulty as the ranks are tied to the goal. However, we make this narrative distinction of intrinsic vs perceived difficulty to highlight *what* we are modifying.

20.3.1 Tuning Intrinsic Difficulty

The intrinsic difficulty is how hard the challenge is based on its design — effectively its mechanical difficulty. For our button mashing challenges this means adjusting the variable components, which in turn affects how much proficiency a player needs to meet the goal. SIBM allows us to tune via time limit, score modifier, or goal. Our model defines the relationship between goal (and by proxy score modifier) and abilities.

By Goal. By reducing the SIBM goal to 60 (the previous normative finger pressing value [32]), we see the Finger Pressing requirement is reduced from 73.78 to 48.85 presses/10 seconds (Fig. 20.11b). Therefore the average player is more likely to surpass that threshold, and their expectations begin to line up with their performance.

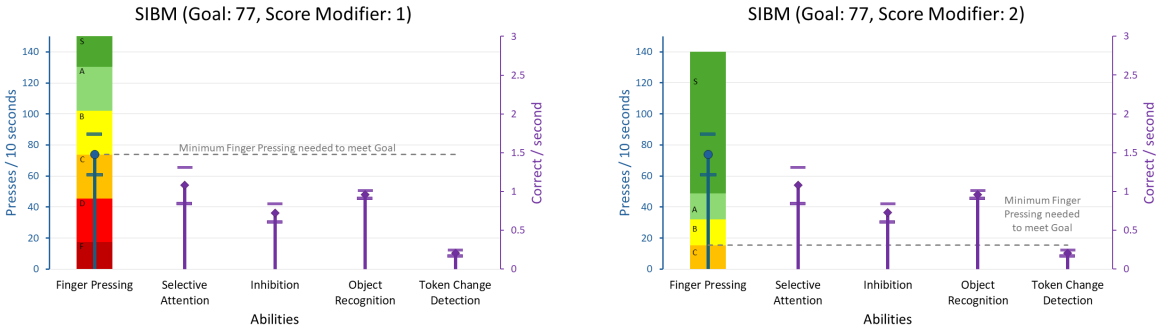


(a) Goal: 77 (b) Goal: 60

Figure 20.11: Comparing SIBM for Average Player based on changing Goal.

By Score Modifier. Since our models are built from regressions on player scores (which incorporate the score modifier), changing the score modifier means using a different model equation. When we double SIBM's baseline score modifier (i.e. underload model), the required Finger Pressing drops from 73.78 to 15.35 presses/10 seconds (Fig. 20.12b). While

this makes the game trivially easy, the ranks align more closely with the average player's expectations of themselves as we saw in Ch. 18.

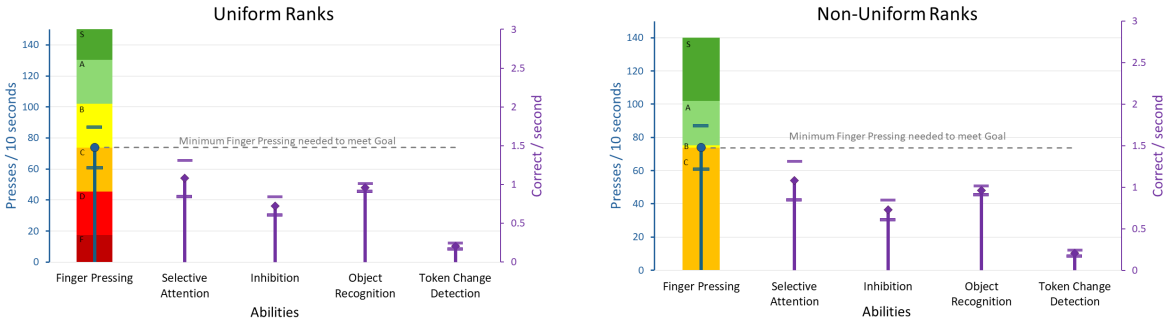


(a) Baseline model (b) Underload model

Figure 20.12: Comparing SIBM for Average Player based on changing Score Modifier.

20.3.2 Tuning for Perceived Difficulty

By playing on player expectations without changing the inherent game design we could influence their perceived difficulty and MX for the challenge. For our button mashing challenges the main way to do this is through changing the thresholds for ranks. Recall the current ranks for our games are uniformly increasing from the goal by 0.25 (e.g. C rank begins at 0.75 *times* Goal, B rank begins at Goal, A rank begins at 1.25 *times* Goal). If instead we use a non-uniform rank, we could skew player's experience as they would have an inflated sense of their performance. Fig. 20.13b shows a version of SIBM at original baseline tuning (Goal: 77, Score Modifier: 1) where the distance between the "success" ranks are really small (A begins at 1.05 × Goal, S begins at 1.25 × Goal), and the "fail" ranks collapse into C-rank (0.25 × Goal).



(a) (b)

Figure 20.13: Comparing SIBM (Goal: 77, Score Modifier: 1) for Average Player between Uniform and Non-Uniform Ranks

20.3.3 Tuning Multiple Variables (Comparing Difficulty Levels)

So far we only discuss changing singular variables. However, we can compare whether difficulty levels have a similar MX when multiple variables are changed by looking at the rank distributions, since the ranks strongly influence the player's perceptions.

Recall our challenge difficulty conditions were achieved by changing what "counts" as a press for each condition (i.e. Score Modifier). We can compare the different models (underload, baseline, overload) perceived experience by comparing their rank distributions given an "effective goal". For button mashing challenges, the "effective goal" is Goal × Score Modifier. For example, consider our goal is 38.5, the effective goal for each model would be: Baseline = 38.5 (Score Modifier: 1), Easy = 77 (Score Modifier: 2), Hard = 19.25 (Score Modifier: 0.5). We look at SIBM and AIBM examples using "effective goals" to compare conditions.

SIBM

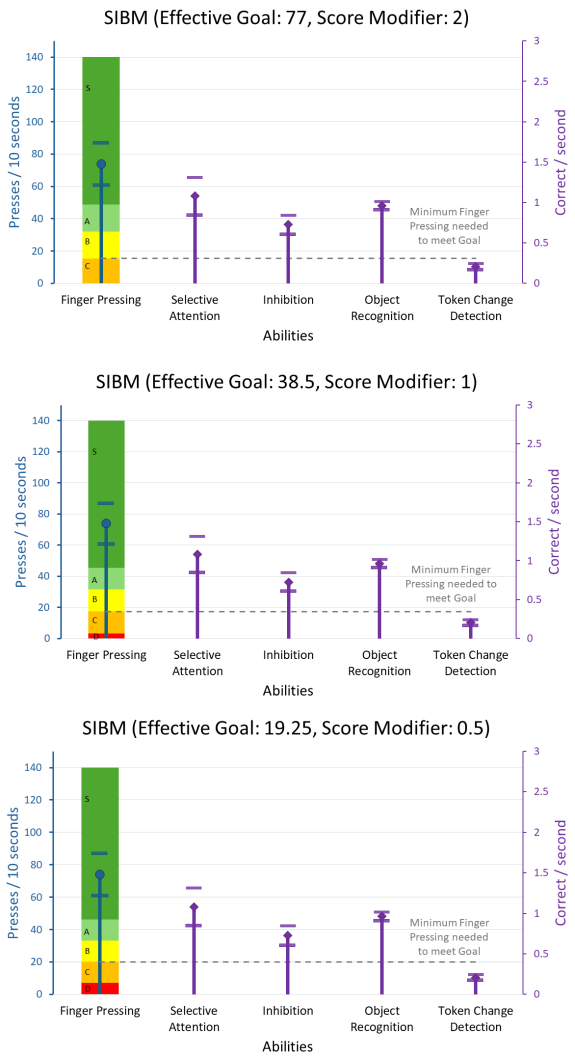


Figure 20.14: Comparing "equivalent" SIBM tunings for Average Player.

Looking at SIBM models for the average player at these "effective goals" (Fig. 20.14) we see the jutsus look almost identical. The average player receives the same rank in each and the minimum required ability is similar (Easy: 15.35, Base: 17.31, Hard: 20.09). While the appearance of potential "D" rank is higher in Baseline and Hard condition, it is really minimal and could be a result of the variance in the condition's equations. This comparison tells us that the models would be perceived the same way if they are tuned similarly. This also shows that, while we experimentally found the different difficulty models (Ch. 17), we could approximate these experiences using "effective goals" and just the Baseline model for SIBM as the scaling is fairly uniform.

AIBM

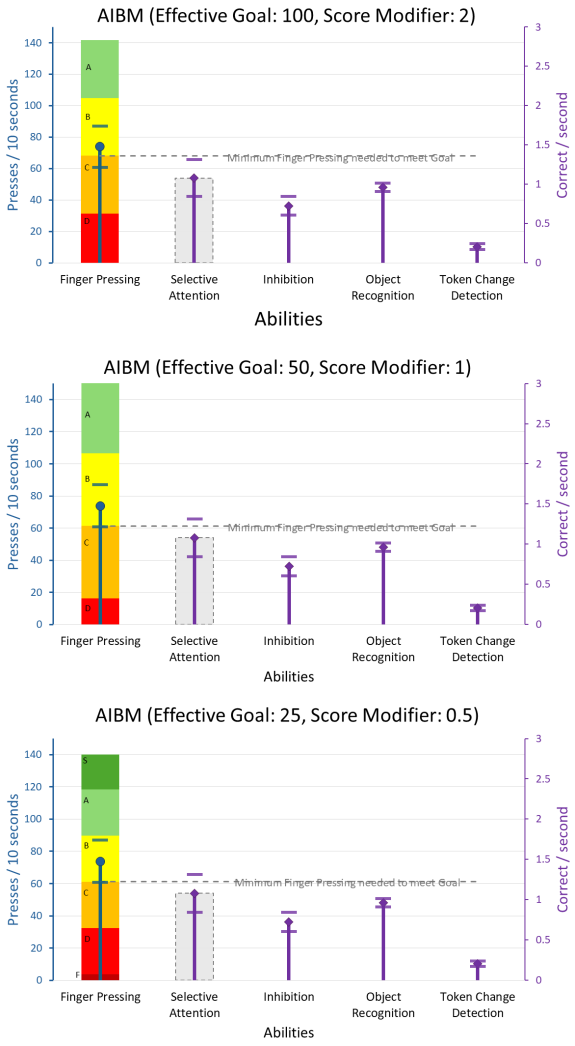


Figure 20.15: Comparing "equivalent" AIBM tunings for Average Player (Threshold: Mean).

In comparison, we found that AIBM did not seem to uniformly scale even though the player scores were effectively uniformly scaled by the score modifier. We compare AIBM models with the same effective goals (i.e. same difficulty) at the Mean (Figs. 20.15) and $-\sigma$ (Figs. 20.16) thresholds. For the average player, these thresholds highlight an average day (Mean - meets the threshold most of the time) and a "good day" ($-\sigma$ - meets and surpasses the threshold all of the time) so we can accurately interpret the finger pressing difficulty. A cursory look shows the scaling does not seem uniform, and the experiences seem oddly dissimilar.

For Mean, the conditions do not seem sufficiently alike; based on the size of the ranks and where the average player's ability lands it almost seems like the Easy condition is harder than the Baseline and Hard conditions in spite of being tuned "the same". For $-\sigma$ we see the experience of the Hard condition does not change from the Mean. But both the Easy and Hard conditions have become more challenging, with players generally dropping to a 50-50 chance of success. Between the three $-\sigma$ conditions, differences in the size of each rank become more apparent.

A notable difference between the Mean and $-\sigma$ thresholds is whether the player's upper selective attention limit is above or below the minimum finger pressing line. This change can be seen in the Baseline conditions, and creates the most significant seeming experiential change. Looking closely, we see that when the minimum finger pressing line is between the player's selective attention thresholds, they seem to have a better rank. We are unclear whether this is meaningful, as the abilities are measured on different scales and this could just be a coincidence in visualizing. However, it may be worth future testing to see how "difficulty" is perceived when the games are tuned so finger pressing requirements fall in that threshold.

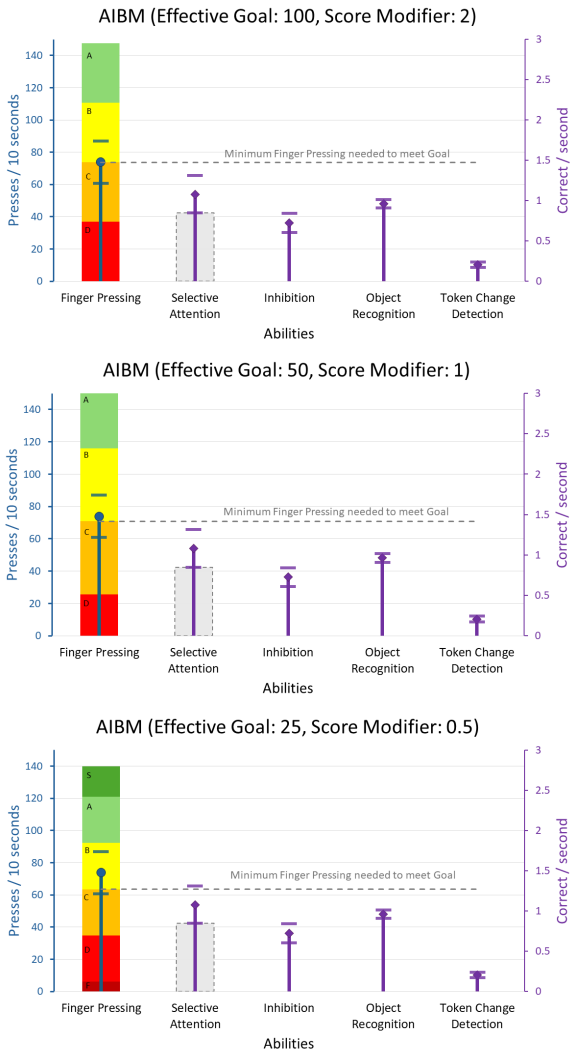


Figure 20.16: Comparing "equivalent" AIBM tunings for Average Player (Threshold: $-\sigma$).

20.4 Comparing Challenges with Jutsu

Beyond assessing mechanical experience and tuning gameplay, we think the jutsu structure can be useful for comparing challenges against each other.

For example, MIBM seems to be intrinsically harder than SIBM. While there ought to be a difference due to increased motor load, the jutsu make it clear that the difference is more perceptible than we originally thought. Playing around with the MIBM jutsu, we see that the challenge becomes more "balanced" (50-50 chance) like SIBM when the goal is 67 (Fig. 20.17, MIBM-67). MIBM-67 has small bands for the F and S-ranks, so players are less likely to catastrophically fail or excessively succeed. Knowing players care about the perception more than the actual likelihoods, this could make them feel like the game is less "fair" and more punishing since they are more stuck in the middle. This further seems to imply that intrinsically MIBMs are harder than SIBMs.

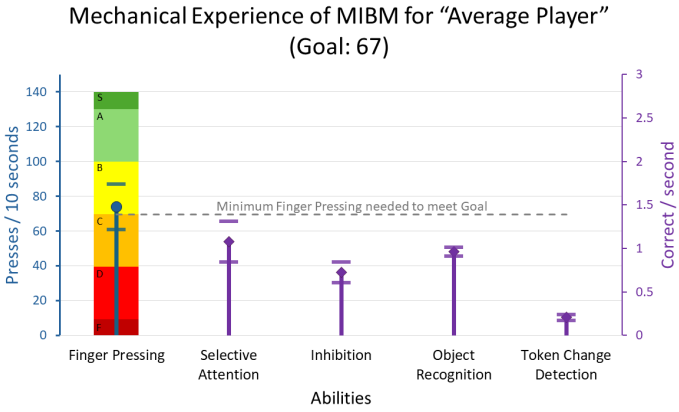


Figure 20.17: MIBM Mechanical Experience for Average Player. Goal: 67

20.5 Moving forward with Jutsu

We present the button mashing jutsu family: single-input button mashing (Fig. 20.6), alternating-input button mashing (Fig. 20.7), and multiple-input button mashing (Fig. 20.9). These jutsu represent the final answers to our research questions, and the first step into a larger body of work on quantifying mechanical experiences.

With our competency profiles and models we can use jutsu to pre-assess the MX and mechanical achievability of different challenge tunings. This would be extremely helpful for coming up with static difficulty levels for these challenges based around an average player profile. This tuning ability could also **help us enforce a specific MX by dynamically tuning these challenges to a player's specific proficiency**.

This **design-time tuning also allows us to speculate about *why* certain values would deliver a "better" experience**. For SIBM, it is possible that player's expectations of high ranks could be reinforced by playing games that were tuned to the old norms and so makes them feel significantly better than the average. If we want to keep this positive feeling we could even try more specific goal refinements to control the ranks (and therefore perceptions) of the players.

Future work should look to both expand the number of jutsu (including testing variants) and undertaking a more qualitative study of the usefulness of jutsu for game designers.

Take home points

From this chapter we learned the following meta-lessons:

- Jutsu successfully visualize mechanical achievability and highlighting potential issues.
- Jutsu can be used to pre-tune a challenge for a particular player profile using the challenge model.
- Jutsu can be used to draw broad comparisons between challenges.
- Jutsu may be useful in trying to "guarantee" a particular mechanical experience through dynamic difficulty tuning.

Chapter 21

Reflecting on The Jutsu Framework

Our thesis goal is to explore player experience (PX) through the lens of experiential types (ExpTypes). ExpTypes reflect the ways a person interacts with a game; they are organized into the Experiential Tetrad (ExperT) to highlight their boundaries and connections to each other. We realise that holistic PX incorporates all ExpTypes, but moves through these experiences from mechanical experience (MX) to socio-cultural. To further explore this idea we choose to focus on thoroughly exploring and modeling MX. We specifically focus on finding basic competency profiles for button mashing challenges and exploring the ways the challenge design affects the limiting ability load and by extension performance. While this is not fully exploring MX, it gives us a starting point in the discussion of this concept.

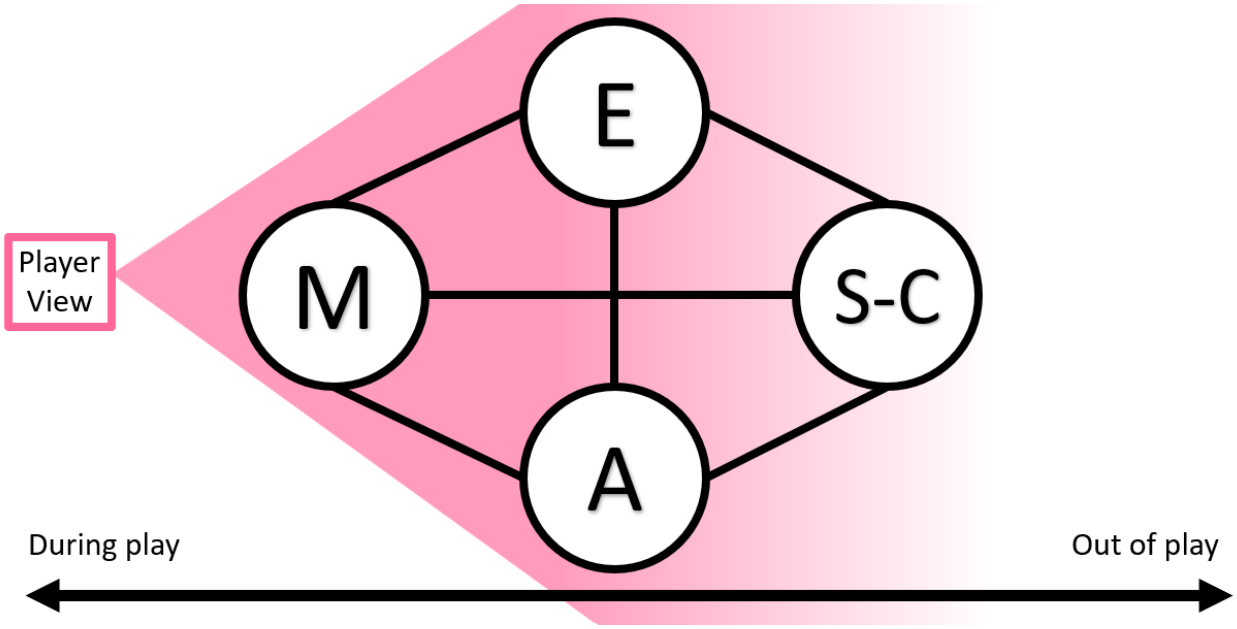


Figure 21.1: Experiential Tetrad from the Player View, outlining four interconnected experiential types: **M**echanical, **E**motional, **A**esthetic, **S**ocio-**C**ultural.

21.1 Thesis in review

We explore MX through the mechanical achievability and mechanical difficulty of button mashing challenges as impacted by their limiting ability. We focus on three questions, and summarize our answers in Table 21.1. These results, while preliminary, are a promising stepping stone in the efforts to quantify MX with jutsus through visualizing competency profiles and how they change with limiting abilities.

Questions	Answers
RQ1: How are cognitive and motor abilities used to interact with various challenges?	• Abilities are a game's "player requirements". • Competency profiles show the abilities used in a challenge, and which are limiting performance.
RQ2: What are the effects of cognitive and motor overloading on the mechanical achievability of a challenge?	• Performance in simple challenges (e.g. SIBM) scales uniformly with load on the limiting ability. • Performance in challenges with multiple important abilities (e.g. AIBM) is more complex. • Perceived experience skews negative when load matches or exceeds player abilities.
RQ3: How can designers use this knowledge?	• Jutsu visualization can help designers tune a challenge's mechanical experience to a player profile at design-time. • Jutsu can visualize how changes to the game mechanics affect mechanical experience.

Table 21.1: Research Question and Our Answers.

21.2 Discussion

While our results show jutsus work at visualizing MX of a challenge, we also found recurring limitations and potential influences on our work.

21.2.1 The Difficulty of Player Modeling

Isolating abilities. Human abilities are not mutually exclusive: abilities overlap in function, and are used in tandem to produce results for any action. Any measurement we make on some quantifiable output, like correct responses or button counts, inherently includes the motor response ability as well as the cognitive abilities we are trying to measure. Our work mirrors cognitive psychology tests and counts the "results" as a score for just the targeted ability. However, this effectively adds a level of uncertainty into the measurements and the resulting models. While this is fine for a preliminary attempt at this work, we wonder if a more robust player profile is feasible. Having more redundant measures in the battery

could allow us to get a weighted-composite score for each individual ability. However, it is unclear how different ability measures should be combined, and it is unclear if it would be advantageous enough to offset the extra fatigue and length of the battery.

Tuning ability measurements. Even if abilities could be easily isolated or combined, tuning the measurement games to get an accurate read of the player's ability is difficult. We tuned the Ability Battery games based on play testing with a variety of volunteer "gamers". These gamers gave us feedback on difficulty and playability. While we knew that these playtests would be skewed towards these gamers' experience, we did not anticipate that their responses may actually have the games be skewed too easy. We see this happen with the Cake game, where most participants got all of the trials correct showing that it was tuned too easy. At the moment, we have not considered this problematic because all of our players went through the same battery and so at least their scores could be compared to each other. However, a more accurate ability measurement would require more tuning to find the level where player abilities begin to fail.

Human variability (Fatigue and Experience). Beyond the initial design difficulties, we consider how player performance in the measured games are different based on environmental and personal factors. At the most basic level, we have no control over whether players are fatigued or distracted when undergoing the modeling process. This could create issues in finding the player's MX since we do not have a good read of their abilities. A more complex, but significant difference, is the variability in player gaming experience and literacy. We noted multiple times that gamers approached basic tasks like button mashing with different strategies based on their gaming literacy and history. Since the ability measurements are similarly perceived as mini-games players applied their same strategies during their assessment. While this gives us a more accurate understanding of what their abilities are in a gaming context, it inadvertently seems to capture strategy/experience as a latent variable. At a larger level this may not be a problem, as a person's performance of an ability cannot be divorced from that kind of practice and history. However, this could mean that when doing future research into jutsus for different challenge types we should seek out more homogeneous groups to get more insight into how different levels of experience might result in different competency profiles.

21.2.2 Perceiving Experiences

Game literacy on challenge modeling. We rely heavily on our own gaming knowledge when reading the challenges we use as a base for our competency profiles. Our personal gaming literacy, and gaming context greatly affects this. The abilities we identify as part of the challenge are directly related to the strategy we use to play the games. While we try to mitigate this effect by playing with particular intent to observe playing using various strategies, it is impossible to guarantee our method is unbiased from our knowledge. Other people using our method for developing competency profiles may therefore end up with somewhat different abilities because of their biases. We think this could become more apparent for complex and cognitively-focused challenges which allow for vastly different strategies/approaches.

Gaming trends. We also consider how gaming trends affect player exposure to different challenge types. As different genres become more popular across different times, entire types of challenges may be erased from the gaming landscape and so player knowledge of it will be varied. We see this in our study on perceived experiences, as players were unsure of how to approach challenges that used to be common in the early 2000s (button mashing).

21.3 Future Work (Ways We Can Keep Learning)

We categorize the different avenues for future work as either improvements to existing work, next steps for MX research, and furthering the theory. We end this section by outlining ways that a jutsu framework could expand the usefulness and accessibility of our work.

21.3.1 Immediate Improvements

Improved measurements. This involves adding redundancy, improving the existing games tuning and design, and creating new games for other abilities. As previously discussed, redundancy and improved existing measurements would allow for a more robust player profile and allow for finding more nuanced relationships between the abilities and gameplay. Fleshing out the battery with games for the rest of the cognitive and motor abilities will also give us the ability to check whether any of those better explain performance, or are latent in our existing games.

Dynamic Difficulty in Button Mashing. At static difficulty levels we end up with players who have different degrees of failure or success at the games. Having the custom button mashing games interface with the Ability Battery would allow us to scale the ranks and difficulty based on the individual player. Thus guaranteeing that players "fail" or "succeed" by the same relative degree in the loading tests. It would also allow us to control for perceived experience since we could make all players feel like their performance met expectations. Work in this direction could create the bedrock for a larger player profile application programming interface (API). With an API for reading ability data during gameplay startup, we could envision a future where these profiling tools are built into the base console experience and all games could read those values to configure themselves for a specific player by default.

Replicating studies with different groups or simulated players. It would be useful to re-run these studies both as-is and with different contexts and groups to see whether the competency profiles are stable. This would add credibility to the larger jutsu concept, and could be incorporated as part of testing jutsu variants. Another potential avenue is to use existing data to create a simulated player which could run the button mashing games multiple times under different conditions. This automated player could give us insight into running the studies thousands of times with various levels of noise in the data to see how stable profiles are, and if there are any other insights we may get from large scale testing.

Experimenting with Jutsu. The usefulness of our specific jutsu is that they only exist for one context. Immediate research following our same methodology, but changing challenge

context variables like the controller, would help us see whether competency profiles change drastically with context or if (when ordered by importance) the shape is the same. This would also add to our ability to construct a jutsu framework as we could explore experiential differences between controllers. However, this kind of variant experimentation would require improvements to the ability battery and minor changes to the custom games.

21.3.2 Next Steps for Mechanical Experience

Extending Cognitive and Motor Models. Our current player model limits itself to motor abilities related to modern controllers, and lower-order cognitive abilities. Future work should look at how to reasonably expand this model. Motor abilities could be reasonably expanded by future reviews of hardware to see whether new abilities are being used. Cognitive abilities are more difficult as we would encourage future work to look into higher-order abilities like executive functioning and problem-solving. These abilities are generally viewed as a sets of interrelated sub-abilities, but their processes are less clear. Work that dives into ways to model these for gameplay would be a significant undertaking, but opens up an ability to examine larger challenge instances that have more open-ended goals or complex mechanics.

Exploring More Challenges Expanding our work requires examining more atomic challenges and developing their jutsu. We suggest future work starts with ***timing challenges*** as the next logical step. These are challenges that seem to require cognitive and motor abilities in similar amounts, and so would give us more insight into the compensatory effects of different abilities and whether limiting abilities change. To this end we offer a preliminary analysis of Balloon Burst [193, 194, 323] in App. C.

21.3.3 Furthering the Theory

Studying other experiences. As we note in our broader contributions, this work could extend into studying different types of experiences or even the relationship between experiences. We think an interesting avenue would be to explore how challenges (or their equivalent) exist in different experience contexts. We would like to get to a stage where we have both a MX and socio-cultural experience (SX) model for a more complex challenge like a level in Hitman. We would want to explore how models of challenges at different levels come together to describe the more holistic PX that other work has focused on studying.

Incorporating Player Context. Currently our understanding of player profiles and experience nominally acknowledges gaming history, literacy, and experiences as affecting experience. A more robust look at how to evaluate a player's gaming context and integrate it into our measures of MX may help give us a deeper understanding as to why a player may score poorly in a measure but perform exceptionally well in the game context.

21.3.4 Imagining a Jutsu Framework

So far we have considered individual jutsu for a small set of challenges tailored to limiting abilities. While this has served as a proof of concept, we know that games are generally made of many interacting atomic challenges. Envisioning the possibility of a large set of jutsus for all of the atomic challenges, we believe a meta-structure that organizes and relates them would make jutsus more useful. We discuss two elements of a larger framework that could be useful: the relationship between jutsus and organization of the jutsu set.

Jutsu Relations

While individual jutsu show us the MX of a challenge, their relationships with other jutsus paint a larger picture about types of experiences in games. We consider three types of jutsu relationships worth discussing: jutsu families, jutsu variants, and complementary/discordant jutsu.

Jutsu Families group together challenges with similar mechanics and competency profiles, like our button mashing challenges. These groupings could be helpful for shortlisting challenges that deliver a similar "feel" so designers can see options that they may not have previously considered.

Jutsu Variants are a version of a jutsu given a different context (e.g. interaction type, number of players, etc). Variants would show how the overall shape of the competency profile does not change when only the motor ability changes. For example, our SIBM jutsu currently uses a button as input and so may work for keyboard and controller schemes. However, if we were to implement SIBM where the "button" was on a dance mat (like Dance Dance Revolution [237]), the new jutsu would rely on leg moving and foot pressing resulting in a similar but different competency profile. A single jutsu may have several related variants.

Variants could help designers quantify the impact of different controllers on MX. A game tuned to be played on a controller (i.e. finger pressing) would have to undergo different tuning for a dance-mat (i.e. foot pressing) because of differences in the movement. In turn this could improve decision making for accessible control schemes. Consider certain players may have difficulty with the fine motor skills of a standard controller, but can play games well with different types of joysticks and switches. With jutsu variants, a designer could see how the control schemes are different as well as how they would need to be tuned to different demographics.

Complementary and Discordant Jutsu express the compatibility of two challenge's MX. Two jutsus are *complementary* if their challenges competency profiles are not conflicting (i.e they have different limiting abilities, and the other required abilities do not exceed the player ability capacity). Conflicting jutsus are *discordant*. As gameplay is concurrently operating atomic challenges, knowing challenge compatibility is important for tuning. Complementary relationships mean challenges are possible to do at the same time; discordant relationships mean challenges will overload the player by design. Future work could explore how the overall MX looks when stacking atomic challenges together.

Organizing Jutsu

With a large set of jutsu, we want to organize them in a way that makes both their individual relationships and the overall state of MX easy to see. Any organization should also be useful for various users to navigate and expand upon. We consider two organizations for the set of jutsu: challenge-based and ability-based.

Challenge-based organization maps out jutsus based on jutsu families (Fig. 21.2) so we can navigate through the jutsus based on the gameplay descriptions. This organization makes finding related challenges easy, and so could help designers find similar experiences. We think this organization can help designers in the beginning stages of their work. If they have a particular challenge that is central to their game, knowing the complementary and discordant relationships with that particular challenge would help find other challenge types to add in the larger game.

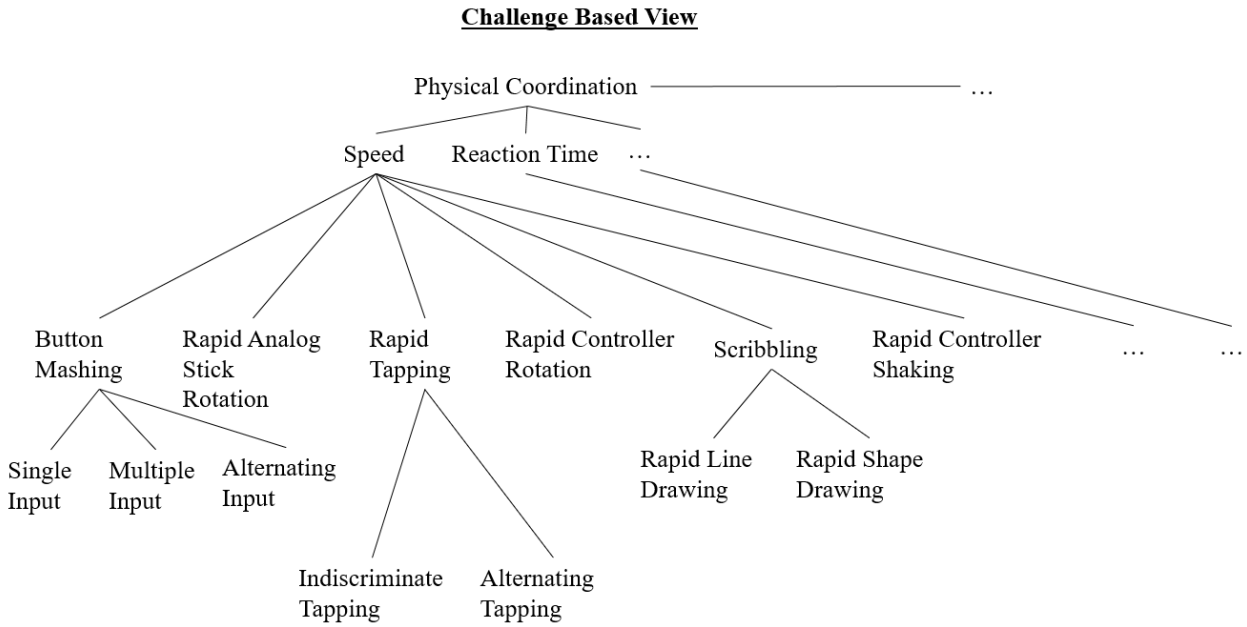


Figure 21.2: Hierarchical challenge based organization of jutsu.

Ability-based organization maps out jutsus based on their limiting abilities (Fig. 21.3) so we can navigate based on targeted abilities. This organization highlights how seemingly unrelated challenges rely on similar skills. This organization may be useful to researchers and medical practitioners who communicate in terms of human abilities. Consider a research team working with children with cerebral palsy who may have trouble with motor function; they already understand the needs and limitations of these children and so can find gameplay that works with their specific requirements. This organization may also help designers tailor games to particular audiences with ability-related concerns. This organization could also be used to find novel gameplay development areas by identifying which abilities are under represented in challenges.

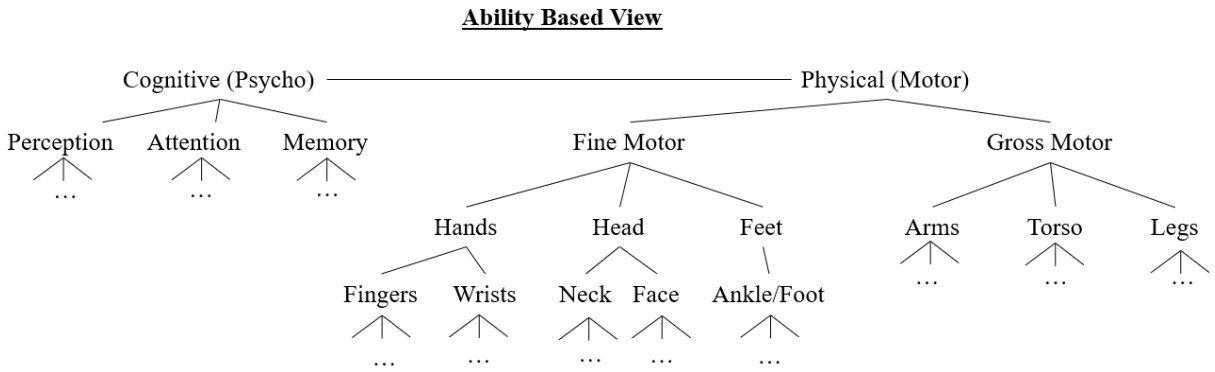


Figure 21.3: Hierarchical ability based organization of jutsu.

21.4 Final Remarks

By starting to explore MX with jutsus we have crossed through many academic domains. In a way, this thesis reflects the practice of game design as an interdisciplinary marriage of science, engineering, and art. The results we have achieved set up a significant body of future research that could impact the way we study game-related experiences. The larger theoretical framework we construct in ExperT opens even more possibilities to the ways that marginalized experiences or intersecting experiences can be studied. So for all our readers who want to continue this quest...

Bibliography

[2] Edwin A. Fleishman, Marilyn K. Quaintance, and Laurie A. Broedling. *Taxonomies of Human Performance: The Description of Human Tasks*. San Diego, CA, USA: Academic Press, Jan. 1984.

[3] Espen Aarseth and Paweł Grabarczyk. "An Ontological Meta-Model for Game Research". In: *DiGRA &#3918 - Proceedings of the 2018 DiGRA International Conference: The Game is the Message*. DiGRA, July 2018. URL: http://www.digra.org/wp-content/uploads/digital-library/DIGRA_2018_paper_247_rev.pdf.

[4] Vero Vanden Abeele et al. "Development and validation of the player experience inventory: A scale to measure player experiences at the level of functional and psychosocial consequences". In: *International Journal of Human-Computer Studies* 135 (2020), p. 102370. ISSN: 1071-5819. DOI: https://doi.org/10.1016/j.ijhcs.2019.102370. URL: https://www.sciencedirect.com/science/article/pii/S1071581919301302.

[5] Johnny Accot and Shumin Zhai. "Beyond Fitts' law: models for trajectory-based HCI tasks". In: *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems*. 1997, pp. 295–302.

[6] Johnny Accot and Shumin Zhai. "Scale Effects in Steering Law Tasks". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '01. Seattle, Washington, USA: Association for Computing Machinery, 2001, 1–8. ISBN: 1581133278. DOI: 10.1145/365024.365027. URL: https://doi.org/10.1145/365024.365027.

[7] Ernest Adams. *Fundamentals of Game Design*. 2nd ed. Berkley, CA, USA: New Riders, 2010.

[8] Victoria K Aldridge, Terence M Dovey, and Angie Wade. "Assessing test-retest reliability of psychological measures". In: *European Psychologist* (2017).

[11] Craig G Anderson. "Hits, quits, and retries-player response to failure in a challenging video game". In: *Proceedings of the 15th International Conference on the Foundations of Digital Games*. 2020, pp. 1–7.

[12] Craig G Anderson, Kathryn Campbell, and Constance Steinkuehler. "Building persistence through failure: the role of challenge in video games". In: *Proceedings of the 14th International Conference on the Foundations of Digital Games*. 2019, pp. 1–6.

[13] John R Anderson, Michael Matessa, and Christian Lebiere. "ACT-R: A theory of higher level cognition and its relation to visual attention". In: *Human–Computer Interaction* 12.4 (1997), pp. 439–462.

[14] Tom Apperley and Catherine Beavis. "A model for critical games literacy". In: *E-learning and Digital Media* 10.1 (2013), pp. 1–12.

[20] Ahmad Azadvar and Alessandro Canossa. "UPEQ: ubisoft perceived experience questionnaire: a self-determination evaluation tool for video games". In: *Proceedings of the 13th International Conference on the Foundations of Digital Games*. FDG '18. Malmö, Sweden: Association for Computing Machinery, 2018. ISBN: 9781450365710. DOI: 10.1145/3235765.3235780. URL: https://doi.org/10.1145/3235765.3235780.

[21] Alan Baddeley. "Exploring the central executive". In: *The Quarterly Journal of Experimental Psychology Section A* 49.1 (1996), pp. 5–28.

[22] Alan D Baddeley. *Human memory: Theory and practice*. Psychology Press, 1997.

[23] Alan D Baddeley, Richard J Allen, and Graham J Hitch. "Binding in visual working memory: The role of the episodic buffer". In: *Neuropsychologia* 49.6 (2011), pp. 1393–1400.

[24] Alan D Baddeley and Graham Hitch. "Working memory". In: *Psychology of learning and motivation* 8 (1974), pp. 47–89.

[25] Alan D Baddeley and Graham J Hitch. "Developments in the concept of working memory." In: *Neuropsychology* 8.4 (1994), p. 485.

[26] Alan D Baddeley, Neil Thomson, and Mary Buchanan. "Word length and the structure of short-term memory". In: *Journal of verbal learning and verbal behavior* 14.6 (1975), pp. 575–589.

[27] Ravin Balakrishnan and I Scott MacKenzie. "Performance differences in the fingers, wrist, and forearm in computer input control". In: *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems*. 1997, pp. 303–310.

[29] Deanna M Barch et al. "CNTRICS final task selection: working memory". In: *Schizophrenia bulletin* 35.1 (2009), pp. 136–152.

[30] Richard Bartle. "Hearts, clubs, diamonds, spades: Players who suit MUDs". In: *Journal of MUD research* 1.1 (1996), p. 19.

[31] Richard Bartle. "Understanding the limits of theory". In: *Beyond Game Design: Nine Steps to Creating Better Videogames*. Ed. by Chris Bateman. Boston, MA, USA: Course Technology - Cengage Learning, 2009. Chap. 4, pp. 117–133.

[32] Çağatay Barut et al. "Advanced analysis of finger-tapping performance: a preliminary study". In: *Balkan medical journal* 2013.2 (2013), pp. 167–171.

[33] Chris Bateman and Richard Boon. *21st Century Game Design (Game Development Series)*. Charles River Media, Inc., 2005.

[34] Chris Bateman, Rebecca Lowenhaupt, Lennart E Nacke, et al. "Player typology in theory and practice." In: *DiGRA &#3911 - Proceedings of the 2011 DiGRA International Conference: Think Design Play*. DiGRA/Utrecht School of the Arts, Jan. 2011.

[35] Melissa R Beck and Amanda E Van Lamsweerde. "Accessing long-term memory representations during visual change detection". In: *Memory & cognition* 39.3 (2011), pp. 433–446.

[36] Jen Beeston et al. "Accessible player experiences (APX): The players". In: *International conference on computers helping people with special needs*. Springer. 2018, pp. 245–253.

[37] Jocelyn J Bélanger et al. "Driven by fear: the effect of success and failure information on passionate individuals' performance." In: *Journal of personality and social psychology* 104.1 (2013), p. 180.

[39] Nicolas Benguigui, Hubert Ripoll, and Michael P Broderick. "Time-to-contact estimation of accelerated stimuli is based on first-order information." In: *Journal of Experimental Psychology: Human Perception and Performance* 29.6 (2003), p. 1083.

[40] Karl Bergström, Staffan Björk, and Sus Lundgren. "Exploring aesthetical gameplay design patterns: camaraderie in four games". In: *Proceedings of the 14th International Academic MindTrek Conference: Envisioning Future Media Environments*. 2010, pp. 17–24.

[41] Kelly Bergstrom. "EVE Online is not for everyone: Exceptionalism in online gaming cultures". In: *Human Technology* 15.3 (2019), pp. 304–325.

[42] Manuela Berlingeri et al. "Anatomy of the episodic buffer: a voxel-based morphometry study in patients with dementia". In: *Behavioural neurology* 19.1-2 (2008), pp. 29–34.

[43] Irving Biederman. "Recognition-by-components: a theory of human image understanding." In: *Psychological review* 94.2 (1987), p. 115.

[44] Frank Biocca, Chad Harms, and Judee K Burgoon. "Toward a more robust theory and measure of social presence: Review and suggested criteria". In: *Presence: Teleoperators & virtual environments* 12.5 (2003), pp. 456–480.

[45] Nicolas Bisson and Simon Grondin. "Time estimates of internet surfing and video gaming". In: *Timing & Time Perception* 1.1 (2013), pp. 39–64.

[46] Jim Bizzocchi and Joshua Tanenbaum. "Well read: Applying close reading techniques to gameplay experiences". In: *Well played 3.0: Video games, value and meaning* 3 (2011), pp. 289–316.

[47] Staffan Bjork and Jussi Holopainen. *Patterns in game design (game development series)*. Needham Heights, MA, USA: Charles River Media, 2004.

[48] Tim Bogenn and Rick Barba. *Rockstar Games presents L.A. Noire*. DK/Brady Games, 2011.

[49] Ian Bogost. *The rhetoric of video games*. MacArthur Foundation Digital Media and Learning Initiative, 2008.

[50] Julia Ayumi Bopp et al. "Exploring emotional attachment to game characters". In: *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. 2019, pp. 313–324.

[51] João Bosco Borges et al. "Player experience evaluation: a brief panorama of instruments and research opportunities". In: *Journal on Interactive Systems* 11.1 (2020), pp. 74–91.

[52] Patrick A Bourke. "A general factor involved in dual task performance decrement". In: *The Quarterly Journal of Experimental Psychology: Section A* 49.3 (1996), pp. 525–545.

[53] Jason T. Bowey, Max V. Birk, and Regan L. Mandryk. "Manipulating Leaderboards to Induce Player Experience". In: *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play*. CHI PLAY '15. London, United Kingdom: Association for Computing Machinery, 2015, 115–120. ISBN: 9781450334662. DOI: 10.1145/2793107.2793138. URL: https://doi.org/10.1145/2793107.2793138.

[54] Nicholas D Bowman and Ron Tamborini. "Task demand and mood repair: The intervention potential of computer games". In: *New Media & Society* 14.8 (2012), pp. 1339–1357.

[55] Nicholas David Bowman, Joseph Wasserman, and Jaime Banks. "Development of the video game demand scale". In: *Video games*. Routledge, 2018, pp. 208–233.

[56] Elizabeth A. Boyle et al. "Engagement in digital entertainment games: A systematic review". In: *Computers in Human Behavior* 28.3 (2012), pp. 771–780. ISSN: 0747-5632. DOI: https://doi.org/10.1016/j.chb.2011.11.020. URL: https://www.sciencedirect.com/science/article/pii/S0747563211002640.

[57] Nyla R Branscombe et al. "The context and content of social identity threat". In: *Social identity: Context, commitment, content* (1999), pp. 35–58.

[58] Brenda Brathwaite and Ian Schreiber. *Challenges for game designers*. Course Technology/Cengage Learning Boston, Massachusetts, 2009.

[59] Virginia Braun and Victoria Clarke. *Thematic analysis*. American Psychological Association, 2012.

[60] Noel Brett and Sasha Soraine. "Pandemic Gaming and Wholesome Philosophy: How New Players Reimaged Gaming Practices". In: *Gaming and Gamers in Times of Pandemic*. Ed. by Piotr Siuda. Ed. by Jakub Majewski. Ed. by Krysztof Chmielewski. Bloomsbury Academic. Chap. 8, pp. 179–200. ISBN: 9798765110232.

[61] Johannes Breuer, Michael Scharkow, and Thorsten Quandt. "Tunnel Vision or Desensitization?" In: *Journal of Media Psychology* (2014).

[62] Katharine C Briggs. *Myers-Briggs type indicator*. Consulting Psychologists Press Palo Alto, CA, 1976.

[63] Jill M Brink and Joan M McDowd. "Aging and selective attention: An issue of complexity or multiple mechanisms?" In: *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences* 54.1 (1999), P30–P33.

[64] Kenneth H Britten and Richard JA van Wezel. "Electrical microstimulation of cortical area MST biases heading perception in monkeys". In: *Nature neuroscience* 1.1 (1998), pp. 59–63.

[65] Donald Eric Broadbent. *Perception and communication*. Elsevier, 2013.

[66] Jeanne H Brockmyer et al. "The development of the Game Engagement Questionnaire: A measure of engagement in video game-playing". In: *Journal of experimental social psychology* 45.4 (2009), pp. 624–634.

[67] Emily Brown and Paul Cairns. "A Grounded Investigation of Game Immersion". In: *CHI '04 Extended Abstracts on Human Factors in Computing Systems*. CHI EA '04. Vienna, Austria: Association for Computing Machinery, 2004, 1297–1300. ISBN: 1581137036. DOI: 10.1145/985921.986048. URL: https://doi.org/10.1145/985921.986048.

[68] Mark Brown and Sky LaRell Anderson. "Designing for Disability: Evaluating the State of Accessibility Design in Video Games". In: *Games and Culture* (Oct. 2020). DOI: 10.1177/1555412020971500. eprint: https://doi.org/10.1177/1555412020971500. URL: https://doi.org/10.1177/1555412020971500.

[70] Sara Bunian et al. "Modeling individual differences in game behavior using HMM". In: *Thirteenth Artificial Intelligence and Interactive Digital Entertainment Conference*. 2017.

[71] Michael Buro and Timothy M Furtak. "RTS games and real-time AI research". In: *Proceedings of the Behavior Representation in Modeling and Simulation Conference (BRIMS)*. Vol. 6370. 2004.

[72] Judith Butler. "Performativity, precarity and sexual politics." In: *AIBR. Revista de Antropología Iberoamericana* 4.3 (2009).

[73] Yuanzhe (Michael) Cai. *Electronic Gaming in the Digital Home*. Sept. 6. URL: http://www.parksassociates.com/research/reports/tocs/2006/multi-gaming.htm.

[74] E.H. Calvillo-Gamez, P. Cairns, and A. Cox. "Assessing the Core Elements of the Gaming Experience". In: *Evaluating User Experience in Games*. Ed. by Regina Bernhaupt. Vol. 1. Human Computer Interaction Series. London, United Kingdom: Springer, 2010. Chap. 3, pp. 47–71.

[78] Stuart K Card, Thomas P Moran, and Allen Newell. "Computer text-editing: An information-processing analysis of a routine cognitive skill". In: *Cognitive psychology* 12.1 (1980), pp. 32–74.

[79] Stuart K Card, Thomas P Moran, and Allen Newell. "The keystroke-level model for user performance time with interactive systems". In: *Communications of the ACM* 23.7 (1980), pp. 396–410.

[80] Stuart K. Card, Allen Newell, and Thomas P. Moran. *The Psychology of Human-Computer Interaction*. Hillsdale, NJ, USA: L. Erlbaum Associates Inc., 1983. ISBN: 0898592437.

[81] Kevin D. Carlson and Andrew O. Herdman. "Understanding the Impact of Convergent Validity on Research Results". In: *Organizational Research Methods* 15.1 (2012), pp. 17–32. DOI: 10.1177/1094428110392383. eprint: https://doi.org/10.1177/1094428110392383. URL: https://doi.org/10.1177/1094428110392383.

[82] Dominic A Carone. *E. Strauss, EMS Sherman, & O. Spreen, A Compendium of Neuropsychological Tests: Administration, Norms, and Commentary: A Review of:", Oxford University Press, New York, 2006."*. 2007.

[83] Loïc Caroux. "Presence in video games: A systematic review and meta-analysis of the effects of game design choices". In: *Applied Ergonomics* 107 (2023), p. 103936.

[84] John M. Carroll and John C. Thomas. "FUN". In: *SIGCHI Bull.* 19.3 (Jan. 1988), 21–24. ISSN: 0736-6906. DOI: 10.1145/49108.1045604. URL: https://doi.org/10.1145/49108.1045604.

[85] Laird S Cermak et al. "The perceptual priming phenomenon in amnesia". In: *Neuropsychologia* 23.5 (1985), pp. 615–622.

[86] Kathy Charmaz. *Constructing grounded theory: A practical guide through qualitative analysis*. sage, 2006.

[87] Jenova Chen. "Flow in Games (and Everything else)". In: *Commun. ACM* 50.4 (Apr. 2007), pp. 31–34. ISSN: 0001-0782. DOI: 10.1145/1232743.1232769. URL: http://doi.acm.org/10.1145/1232743.1232769.

[88] EC Cherry. "Some experiments on the recognition of speech, with one and with two ears." In: *Journal of the Acoustical Society of America* 25 (1953), pp. 975–979.

[89] Dongkyu Choi et al. "A Believable Agent for First-Person Shooter Games." In: *AIIDE*. 2007, pp. 71–73.

[91] Tom Cole, Paul Cairns, and Marco Gillies. "Emotional and Functional Challenge in Core and Avant-Garde Games". In: *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play*. CHI PLAY '15. London, United Kingdom: Association for Computing Machinery, 2015, 121–126. ISBN: 9781450334662. DOI: 10.1145/2793107.2793147. URL: https://doi.org/10.1145/2793107.2793147.

[92] Fabienne Collette et al. "Exploring the unity and diversity of the neural substrates of executive functioning". In: *Human brain mapping* 25.4 (2005), pp. 409–423.

[94] Mia Consalvo and Christopher A Paul. *Real games: What's legitimate and what's not in contemporary videogames*. MIT Press, 2019.

[97] Paul T Costa and Robert R McCrae. "A five-factor theory of personality". In: *The five-factor model of personality: Theoretical perspectives* 2 (1999), pp. 51–87.

[98] Nelson Cowan. "The magical mystery four: How is working memory capacity limited, and why?" In: *Current directions in psychological science* 19.1 (2010), pp. 51–57.

[99] Nelson Cowan et al. "Models of verbal working memory capacity: What does it take to make them work?" In: *Psychological review* 119.3 (2012), p. 480.

[100] John W Creswell and Vicki L Plano Clark. *Designing and conducting mixed methods research*. Sage publications, 2017.

[101] John W Creswell and J David Creswell. *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications, 2017.

[103] Mihaly Csikszentmihalyi. *Finding flow: The psychology of engagement with everyday life*. New York, NY, USA: Basic Books, 1997.

[104] Roe Curie and Mitchell Alex. "&ldquoIs This Really Happening?&rdquo: Game Mechanics as Unreliable Narrator". In: *DiGRA &#3919 - Proceedings of the 2019 DiGRA International Conference: Game, Play and the Emerging Ludo-Mix*. DiGRA, Aug. 2019. URL: http://www.digra.org/wp-content/uploads/digital-library/DiGRA_2019_paper_201-min.pdf.

[105] Jodi L Davenport and Mary C Potter. "Scene consistency in object and background perception". In: *Psychological Science* 15.8 (2004), pp. 559–564.

[106] Frederik De Grove, Cédric Courtois, and Jan Van Looy. "How to be a gamer! Exploring personal and social indicators of gamer identity". In: *Journal of Computer-Mediated Communication* 20.3 (2015), pp. 346–361.

[107] Michael Anthony DeAnda and Carly A Kocurek. "Game design as technical communication: Articulating game design through textbooks". In: *Technical Communication Quarterly* 25.3 (2016), pp. 202–210.

[108] Alena Denisova, Christian Guckelsberger, and David Zendle. "Challenge in digital games: Towards developing a measurement tool". In: *Proceedings of the 2017 chi conference extended abstracts on human factors in computing systems*. 2017, pp. 2511–2519.

[109] Alena Denisova, A. Imran Nordin, and Paul Cairns. "The Convergence of Player Experience Questionnaires". In: *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play*. CHI PLAY '16. Austin, Texas, USA: Association for Computing Machinery, 2016, 33–37. ISBN: 9781450344562. DOI: 10.1145/2967934.2968095. URL: https://doi.org/10.1145/2967934.2968095.

[110] Alena Denisova et al. "Measuring perceived challenge in digital games: Development & validation of the challenge originating from recent gameplay interaction scale (CORGIS)". In: *International Journal of Human-Computer Studies* 137 (2020), p. 102383.

[111] Michael D'errico. "Worlds of sound: indie games, proceduralism, and the aesthetics of emergence". In: *Music, sound, and the moving image* 9.2 (2016), pp. 191–206.

[112] Aditya Deshbandhu. "Player perspectives: what it means to be a gamer". In: *Press Start* 3.2 (2016), pp. 48–64.

[113] Roberto Dillon. "Teaching games through the AGE framework". In: *IGDA Perspectives Newsletter* 2012 (2012), pp. 1–7.

[114] Roberto Dillon. "The 6-11 Framework: A new methodology for game analysis and design". In: *Proceedings Game-On Asia Conference, Singapore*. 2011, pp. 25–29.

[115] Damien Djaouti et al. "A gameplay definition through videogame classification". In: *International Journal of Computer Games Technology* 2008 (2008), p. 4.

[116] Pawel Dobrowolski et al. "Perceptual, attentional, and executive functioning after real-time strategy video game training: Efficacy and relation to in-game behavior". In: *Journal of Cognitive Enhancement* (2021), pp. 1–14.

[117] A. Drachen, A. Canossa, and G. N. Yannakakis. "Player modeling using self-organization in Tomb Raider: Underworld". In: *2009 IEEE Symposium on Computational Intelligence and Games*. 2009, pp. 1–8.

[118] C. Draheim et al. "Reaction time in differential and developmental research: A review and commentary on the problems and alternatives". In: *Psychological Bulletin* 145.5 (2019), pp. 508–535. DOI: 10.1037/bul0000192.

[119] Wlodzislaw Duch, Richard Jayadi Oentaryo, and Michel Pasquier. "Cognitive Architectures: Where do we go from here?" In: *AGI*. Vol. 171. 2008, pp. 122–136.

[120] Brad Duchaine and Ken Nakayama. "Dissociations of face and object recognition in developmental prosopagnosia". In: *Journal of Cognitive Neuroscience* 17.2 (2005), pp. 249–261.

[121] Daniel Dunne. "Multimodality or Ludo-Narrative Dissonance: Duality of Presentation in Fringe Media". In: *Proceedings of the 2014 Conference on Interactive Entertainment*. IE2014. Newcastle, NSW, Australia: Association for Computing Machinery, 2014, 1–4. ISBN: 9781450327909. DOI: 10.1145/2677758.2677785. URL: https://doi.org/10.1145/2677758.2677785.

[122] Matthew WG Dye, C Shawn Green, and Daphne Bavelier. "The development of attention skills in action video game players". In: *Neuropsychologia* 47.8-9 (2009), pp. 1780–1789.

[123] Adeleh Ebrahimi and Mohammad-R Akbarzadeh-T. "Dynamic difficulty adjustment in games by using an interactive self-organizing architecture". In: *2014 Iranian Conference on Intelligent Systems (ICIS)*. IEEE. 2014, pp. 1–6.

[124] Abdelkader Ennaceur and Jean Delacour. "A new one-trial test for neurobiological studies of memory in rats. 1: Behavioral data". In: *Behavioural brain research* 31.1 (1988), pp. 47–59.

[125] *Ergonomics of human-system interaction*. Chap. ISO Standard 9241-411: Evaluation methods for the design of physical input devices. 62 pp. URL: https://www.iso.org/obp/ui/en/#iso:std:iso:ts:9241:-411:ed-1:v1:en.

[126] Barbara A Eriksen and Charles W Eriksen. "Effects of noise letters upon the identification of a target letter in a nonsearch task". In: *Perception & psychophysics* 16.1 (1974), pp. 143–149.

[127] Laura Ermi and Frans Mäyrä. "Fundamental components of the gameplay experience: Analysing immersion". In: *Worlds in play: International perspectives on digital games research* 37.2 (2005), pp. 37–53.

[129] M.W. Eysenck and M.T. Keane. *Cognitive Psychology A Student's Handbook, 6th Edition*. Taylor & Francis, 2010. ISBN: 9781841695402.

[130] Carlo Fabricatore, Miguel Nussbaum, and Ricardo Rosas. "Playability in action videogames: A qualitative design model". In: *Human-Computer Interaction* 17.4 (2002), pp. 311–368.

[132] Stephen H Fairclough and Louise Venables. "Prediction of subjective states from psychophysiology: A multivariate approach". In: *Biological psychology* 71.1 (2006), pp. 100–110.

[133] Martha J Farah. "Is face recognition 'special'? Evidence from neuropsychology". In: *Behavioural brain research* 76.1 (1996), pp. 181–189.

[134] Martha J Farah, Karen L Levinson, and Karen L Klein. "Face perception and within-category discrimination in prosopagnosia". In: *Neuropsychologia* 33.6 (1995), pp. 661–674.

[135] Martha J Farah et al. "What is special about face perception?" In: *Psychological review* 105.3 (1998), p. 482.

[136] John Feil and Marc Scattergood. *Beginning game level design*. Boston, MA, USA: Thomson Course Technology, 2005.

[137] Leah M Feuerstahler et al. "A note on the identification of change detection task models to measure storage capacity and attention in visual working memory". In: *Behavior research methods* 51.3 (2019), pp. 1360–1370.

[138] David T Field, Richard M Wilkie, and John P Wann. "Neural systems in the visual control of steering". In: *Journal of Neuroscience* 27.30 (2007), pp. 8002–8010.

[139] Paul M Fitts. "The information capacity of the human motor system in controlling the amplitude of movement." In: *Journal of experimental psychology* 47.6 (1954), p. 381.

[140] Alex Flint, Alena Denisova, and Nick Bowman. "Comparing Measures of perceived challenge and demand in video games: Exploring the conceptual dimensions of CORGIS and VGDS". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 2023, pp. 1–19.

[141] Friedrich Försterling. *Attribution: An introduction to theories, research and applications*. Psychology Press, 2013.

[142] J. Fraenkel, N. Wallen, and H. Hyun. *How to Design and Evaluate Research in Education*. 11th ed. New York, New York, USA: McGraw Hill, 2023. ISBN: 978-1-265-18481-0.

[143] Stan Franklin et al. "LIDA: A systems-level architecture for cognition, emotion, and learning". In: *IEEE Transactions on Autonomous Mental Development* 6.1 (2014), pp. 19–41.

[144] Naomi P Friedman and Akira Miyake. "The relations among inhibition and interference control functions: a latent-variable analysis." In: *Journal of experimental psychology: General* 133.1 (2004), p. 101.

[145] Tobias Fritsch, Benjamin Voigt, and Jochen Schiller. "Distribution of online hardcore player behavior: (how hardcore are you?)" In: *Proceedings of 5th ACM SIGCOMM workshop on Network and system support for games*. 2006, 16–es.

[146] Julian Frommel, Madison Klarkowski, and Regan L. Mandryk. "The Struggle is Spiel: On Failure and Success in Games". In: *Proceedings of the 16th International Conference on the Foundations of Digital Games*. FDG '21. Montreal, QC, Canada: Association for Computing Machinery, 2021. ISBN: 9781450384223. DOI: 10.1145/3472538.3472565. URL: https://doi.org/10.1145/3472538.3472565.

[147] Tracy Fullerton. *Game design workshop: a playcentric approach to creating innovative games*. CRC press, 2014.

[148] John DE Gabrieli et al. "Double dissociation between memory systems underlying explicit and implicit memory in the human brain". In: *Psychological Science* 6.2 (1995), pp. 76–82.

[149] Kathleen M Galotti. *Cognitive development Infancy through adolescence*. Sage Publications, 2015.

[152] Isabel Gauthier and Michael J Tarr. "Becoming a "Greeble" expert: Exploring mechanisms for face recognition". In: *Vision research* 37.12 (1997), pp. 1673–1682.

[153] Isabel Gauthier and Michael J Tarr. "Unraveling mechanisms for expert object recognition: bridging brain activity and behavior." In: *Journal of Experimental Psychology: Human Perception and Performance* 28.2 (2002), p. 431.

[154] Manuel Gentile and Antonio Lieto. "The role of mental rotation in TetrisTM gameplay: An ACT-R computational cognitive model". In: *Cognitive Systems Research* 73 (2022), pp. 1–11.

[156] Brett Gibson, Edward Wasserman, and Steven J Luck. "Qualitative similarities in the visual short-term memory of pigeons and people". In: *Psychonomic bulletin & review* 18.5 (2011), pp. 979–984.

[157] Jeremy Gibson. *Introduction to Game Design, Prototyping, and Development: From Concept to Playable Game with Unity and C*. Addison-Wesley Professional, 2014, pp. 31–36.

[158] Joshua L Gills et al. "Validation of a digitally delivered visual paired comparison task: reliability and convergent validity with established cognitive tests". In: *Geroscience* 41 (2019), pp. 441–454.

[159] Ben Godde and Claudia Voelcker-Rehage. "Cognitive Resources Necessary for Motor Control in Older Adults Are Reduced by Walking and Coordination Training". In: *Frontiers in Human Neuroscience* 11 (2017), p. 156. ISSN: 1662-5161. DOI: 10.3389/fnhum.2017.00156. URL: https://www.frontiersin.org/article/10.3389/fnhum.2017.00156.

[160] Ben Goertzel, Cassio Pennachin, and Nil Geisweiller. "Brief Survey of Cognitive Architectures". In: *Engineering General Intelligence, Part 1*. Springer, 2014, pp. 101–142.

[162] Melvyn A Goodale and A David Milner. "Separate visual pathways for perception and action". In: *Trends in neurosciences* 15.1 (1992), pp. 20–25.

[163] PA Gooding, CL Isaac, and AR Mayes. "Prose recall and amnesia: more implications for the episodic buffer". In: *Neuropsychologia* 43.4 (2005), pp. 583–587.

[164] C Goodwin. "Professional Vision, in "American Anthropologist", 96 (3)". In: *TECNOSCIENZA–3 (1)* 47 (1994).

[165] Daniel Gopher. "Task difficulty, resources, and dual-task performance". In: *Attention and performance VIII* 8 (1980), p. 297.

[166] Peter Graf and Daniel L Schacter. "Implicit and explicit memory for new associations in normal and amnesic subjects." In: *Journal of Experimental Psychology: Learning, memory, and cognition* 11.3 (1985), p. 501.

[167] Peter Graf, Larry R Squire, and George Mandler. "The information that amnesic patients do not forget." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 10.1 (1984), p. 164.

[168] Franklin Graybill and Hariharan Iyer. *Regression Analysis: Concepts and Applications*. Belmont, California, USA: Duxbury Press, 1994.

[169] C Shawn Green and Daphne Bavelier. "Action video game modifies visual selective attention". In: *Nature* 423.6939 (2003), pp. 534–537.

[170] Mark D Griffiths, Mark NO Davies, and Darren Chappell. "Breaking the stereotype: The case of online gaming". In: *CyberPsychology & behavior* 6.1 (2003), pp. 81–91.

[171] Kalanit Grill-Spector and Nancy Kanwisher. "Visual recognition: As soon as you know it is there, you know what it is". In: *Psychological Science* 16.2 (2005), pp. 152–160.

[172] Martin Grunwald et al. "Theta power in the EEG of humans during ongoing processing in a haptic object recognition task". In: *Cognitive Brain Research* 11.1 (2001), pp. 33–37.

[173] E Grace Otto-de Haart, David P Carey, and Alan B Milne. "More thoughts on perceiving and grasping the Muller–Lyer illusion". In: *Neuropsychologia* 37.13 (1999), pp. 1437–1444.

[174] Ivan Hagendoorn. "Some speculative hypotheses about the nature and perception of dance and choreography". In: *Journal of Consciousness Studies* 11.3-4 (2004), pp. 79–110.

[178] Juho Hamari and Janne Tuunanen. "Player types: a meta-synthesis". English. In: *Transactions of the Digital Games Research Association* 1.2 (2014), pp. 29–53. ISSN: 2328-9422. DOI: 10.26503/todigra.v1i2.13.

[179] N Hamilton, W Weimar, and K Luttgens. *Kinesiology: Scientific basis of human motion*. 2012.

[180] William A. Hamilton, Oliver Garretson, and Andruid Kerne. "Streaming on Twitch: Fostering Participatory Communities of Play within Live Mixed Media". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '14. Toronto, Ontario, Canada: Association for Computing Machinery, 2014, 1315–1324. ISBN: 9781450324731. DOI: 10.1145/2556288.2557048. URL: https://doi.org/10.1145/2556288.2557048.

[182] C. Hedge, G. Powell, and P. Sumner. "The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences". In: *Behavioural Research* 50 (June 2018), pp. 1166–1186. DOI: 10.3758/s13428-017-0935-1.

[183] Carrie Heeter et al. "Game design and the challenge-avoiding, self-validator player type". In: *International Journal of Gaming and Computer-Mediated Simulations (IJGCMS)* 1.3 (2009), pp. 53–67.

[184] Mary Hegarty, Priti Shah, and Akira Miyake. "Constraints on using the dual-task methodology to specify the degree of central executive involvement in cognitive tasks". In: *Memory & Cognition* 28.3 (2000), pp. 376–385.

[185] Douglas A. Henderson and Daniel R. Denison. "Stepwise Regression in Social and Psychological Research". In: *Psychological Reports* 64.1 (1989), pp. 251–257. DOI: 10.2466/pr0.1989.64.1.251. eprint: https://doi.org/10.2466/pr0.1989.64.1.251. URL: https://doi.org/10.2466/pr0.1989.64.1.251.

[186] RNA Henson, Neil Burgess, and Christopher D Frith. "Recoding, storage, rehearsal and grouping in verbal short-term memory: an fMRI study". In: *Neuropsychologia* 38.4 (2000), pp. 426–440.

[187] William E Hick. "On the rate of gain of information". In: *Quarterly Journal of experimental psychology* 4.1 (1952), pp. 11–26.

[188] Andrew Hollingworth. "Failures of retrieval and comparison constrain change detection in natural scenes." In: *Journal of Experimental Psychology: Human Perception and Performance* 29.2 (2003), p. 388.

[189] Andrew Hollingworth and John M Henderson. "Accurate visual memory for previously attended objects in natural scenes." In: *Journal of Experimental Psychology: Human Perception and Performance* 28.1 (2002), p. 113.

[190] Andrew Hollingworth and John M Henderson. "Object identification is isolated from scene semantic constraint: Evidence from object type and token discrimination". In: *Acta psychologica* 102.2 (1999), pp. 319–343.

[191] Andrew Hollingworth and John M Henderson. "Sustained change blindness to incremental scene rotation: A dissociation between explicit change detection and visual memory". In: *Perception & Psychophysics* 66.5 (2004), pp. 800–807.

[192] Gerald A. Hudgens et al. "Hand Steadiness: Effects of Sex, Menstrual Phase, Oral Contraceptives, Practice, and Handgun Weight". In: *Human Factors* 30.1 (1988). PMID: 3350527, pp. 51–60. DOI: 10.1177/001872088803000105. eprint: https://doi.org/10.1177/001872088803000105. URL: https://doi.org/10.1177/001872088803000105.

[199] Robin Hunicke and Vernell Chapman. "AI for Dynamic Difficulty Adjustment in Games. 2004". In: *Association for the Advancement of Artificial Intelligence (AAAI)* (2004), pp. 2–5.

[200] Robin Hunicke, Marc LeBlanc, and Robert Zubek. "MDA: A formal approach to game design and game research". In: *Proceedings of the AAAI Workshop on Challenges in Game AI*. Vol. 4. 1. 2004, p. 1722.

[201] Ray Hyman. "Stimulus information as a determinant of reaction time." In: *Journal of experimental psychology* 45.3 (1953), p. 188.

[202] Ioanna Iacovides et al. "Player strategies: Achieving breakthroughs and progressing in single-player and cooperative games". In: *Proceedings of the first ACM SIGCHI annual symposium on Computer-human interaction in play*. 2014, pp. 131–140.

[204] Wijnand IJsselsteijn et al. "Characterising and measuring user experiences in digital games". In: *International conference on advances in computer entertainment technology*. Vol. 2. 2007, p. 27.

[205] Wijnand A IJsselsteijn et al. "Presence: concept, determinants, and measurement". In: *Human vision and electronic imaging V*. Vol. 3959. International Society for Optics and Photonics. 2000, pp. 520–529.

[206] Yavuz Inal and Jo Wake. "An old game, new experience: exploring the effect of players' personal gameplay history on game experience". In: *Universal Access in the Information Society* 22.3 (2023), pp. 757–769.

[209] LS Jakobson et al. "A kinematic analysis of reaching and grasping movements in a patient recovering from optic ataxia". In: *Neuropsychologia* 29.8 (1991), pp. 803–809.

[210] Thomas W James et al. "Ventral occipital lesions impair object recognition but not object-directed grasping: an fMRI study". In: *Brain* 126.11 (2003), pp. 2463–2475.

[211] Tiffany S Jastrzembski and Neil Charness. "The Model Human Processor and the older adult: parameter estimation and validation within a mobile phone task." In: *Journal of Experimental Psychology Applied* 13.4 (2007), p. 224.

[212] Charlene Jennett et al. "Measuring and defining the experience of immersion in games". In: *International Journal of Human-Computer Studies* 66.9 (2008), pp. 641–661. ISSN: 1071-5819. DOI: https://doi.org/10.1016/j.ijhcs.2008.04.004. URL: https://www.sciencedirect.com/science/article/pii/S1071581908000499.

[213] Bonnie E John. "Extensions of GOMS analyses to expert performance requiring perception of dynamic visual and auditory information". In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. 1990, pp. 107–116.

[214] Bonnie E John and Wayne D Gray. "CPM-GOMS: an analysis method for tasks with parallel activities". In: *Conference companion on Human factors in computing systems*. 1995, pp. 393–394.

[215] Bonnie E John, Alonso H Vera, and Allen Newell. "Towards real-time GOMS: a model of expert behaviour in a highly interactive task". In: *Behaviour & Information Technology* 13.4 (1994), pp. 255–267.

[216] John Spinello. Operation. Game [Board game]. Hasbro, Rhode Island, USA. Rhode Island, USA, Mar. 1965.

[217] Daniel Johnson, M. John Gardner, and Ryan Perry. "Validation of two game experience scales: The Player Experience of Need Satisfaction (PENS) and Game Experience Questionnaire (GEQ)". In: *International Journal of Human-Computer Studies* 118 (2018), pp. 38–46. ISSN: 1071-5819. DOI: https://doi.org/10.1016/j.ijhcs.2018.05.003. URL: https://www.sciencedirect.com/science/article/pii/S1071581918302337.

[218] Linderoth Jonas and Bennerstedt Ulrika. "This is not a Door: an Ecological approach to Computer Games". In: *DiGRA &#3907 - Proceedings of the 2007 DiGRA International Conference: Situated Play*. The University of Tokyo, Sept. 2007. ISBN: ISSN 2342-9666. URL: http://www.digra.org/wp-content/uploads/digital-library/07312.51011.pdf.

[219] Jesper Juul. *A casual revolution: Reinventing video games and their players*. MIT press, 2010.

[220] Jesper Juul. *The art of failure: An essay on the pain of playing video games*. MIT press, 2013.

[221] Daniel Kahneman. *Attention and effort*. Vol. 1063. Citeseer, 1973.

[222] Robert L Kane and Gary G Kay. "Computerized assessment in neuropsychology: a review of tests and test batteries". In: *Neuropsychology review* 3.1 (1992), pp. 1–117.

[223] Katelynn A Kapalo et al. "Individual differences in video gaming: Defining hardcore video gamers". In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Vol. 59. 1. SAGE Publications Sage CA: Los Angeles, CA. 2015, pp. 878–881.

[224] Narinder Kapur. "Syndromes of retrograde amnesia: a conceptual and empirical synthesis." In: *Psychological bulletin* 125.6 (1999), p. 800.

[225] Veli-Matti Karhulahti. "Mechanic/Aesthetic Videogame Genres: Adventure and Adventure". In: *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*. MindTrek '11. Tampere, Finland: Association for Computing Machinery, 2011, 71–74. ISBN: 9781450308168. DOI: 10.1145/2181037.2181050. URL: https://doi.org/10.1145/2181037.2181050.

[226] Michael M Kasumovic and Jeffrey H Kuznekoff. "Insights into sexism: Male status and performance moderates female-directed hostile and amicable behaviour". In: *PloS one* 10.7 (2015), e0131613.

[227] Jozsef Katona and Attila Kovari. "The evaluation of bci and pebl-based attention tests". In: *Acta Polytechnica Hungarica* 15.3 (2018), pp. 225–249.

[228] Dominik Kayser, Sebastian Andrea Caesar Perrig, and Florian Brühlmann. "Measuring Players' Experience of Need Satisfaction in Digital Games: An Analysis of the Factor Structure of the UPEQ". In: *Extended Abstracts of the 2021 Annual Symposium on Computer-Human Interaction in Play*. CHI PLAY '21. Virtual Event, Austria: Association for Computing Machinery, 2021, 158–162. ISBN: 9781450383561. DOI: 10.1145/3450337.3483499. URL: https://doi.org/10.1145/3450337.3483499.

[229] David E Kieras and David E Meyer. "An overview of the EPIC architecture for cognition and performance with application to human-computer interaction". In: *Human-computer interaction* 12.4 (1997), pp. 391–438.

[230] David E Kieras, Scott D Wood, and David E Meyer. "Predictive engineering models based on the EPIC architecture for a multimodal high-performance human-computer interaction task". In: *ACM Transactions on Computer-Human Interaction (TOCHI)* 4.3 (1997), pp. 230–275.

[231] David E Kieras et al. *Insights into working memory from the perspective of the EPIC architecture for modeling skilled perceptual-motor and cognitive human performance.* Tech. rep. MICHIGAN UNIV ANN ARBOR DIV OF RESEARCH DEVELOPMENT and ADMINISTRATION, 1998.

[232] Min Gyu Kim and Joohan Kim. "Cross-validation of reliability, convergent and discriminant validity for the problematic online game use scale". In: *Computers in Human Behavior* 26.3 (2010), pp. 389–398. ISSN: 0747-5632. DOI: https://doi.org/10.1016/j.chb.2009.11.010. URL: https://www.sciencedirect.com/science/article/pii/S0747563209001812.

[233] Sergey Kiselev, Kimberly Andrews Espy, and Tiffany Sheffield. "Age-related differences in reaction time task performance in young children". In: *Journal of Experimental Child Psychology* 102.2 (2009), pp. 150–166.

[234] Karl Christoph Klauer and Zengmei Zhao. "Double dissociations in visual and spatial short-term memory." In: *Journal of Experimental Psychology: General* 133.3 (2004), p. 355.

[235] Vojtěch Klézl and Stephen Kelly. "Negativists, enthusiasts and others: a typology of players in free-to-play games". In: *Multimedia Tools and Applications* 82.5 (2023), pp. 7939–7960.

[236] Christoph Klimmt et al. "Experimental evidence for suspense as determinant of video game enjoyment". In: *Cyberpsychology & behavior* 12.1 (2009), pp. 29–31.

[239] Seiki Konishi et al. "No-go dominant brain activity in human inferior prefrontal cortex revealed by functional magnetic resonance imaging". In: *European Journal of Neuroscience* 10.3 (1998), pp. 1209–1213.

[240] T. Koo and M. Li. "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research". In: *Journal of Chiropractic Medicine* 15 (2 June 2016), pp. 155–163. DOI: 10.1016/j.jcm.2016.02.012.

[241] Bruno Kopp, Uwe Mattler, and Fred Rist. "Selective attention and response competition in schizophrenic patients". In: *Psychiatry research* 53.2 (1994), pp. 129–139.

[242] Thomas Kosch et al. "A Survey on Measuring Cognitive Workload in Human-Computer Interaction". In: *ACM Comput. Surv.* 55.13s (July 2023). ISSN: 0360-0300. DOI: 10.1145/3582272. URL: https://doi.org/10.1145/3582272.

[243] Rachel Kowert, Ruth Festl, and Thorsten Quandt. "Unpopular, overweight, and socially inept: Reconsidering the stereotype of online gamers". In: *Cyberpsychology, Behavior, and Social Networking* 17.3 (2014), pp. 141–146.

[244] Jussi Kuittinen et al. "Casual games discussion". In: *Proceedings of the 2007 conference on Future Play*. 2007, pp. 105–112.

[245] F Lacquaniti, M Carrozzo, and NA Borghese. "The role of vision in tuning anticipatory motor responses of the limbs". In: *Multisensory control of movement* (1993), pp. 379–393.

[246] John E Laird. "An exploration into computer games and computer generated forces". In: *Eighth Conference on Computer Generated Forces and Behavior Representation*. Citeseer. 2000.

[247] John E. Laird. "It Knows What You'Re Going to Do: Adding Anticipation to a Quakebot". In: *Proceedings of the Fifth International Conference on Autonomous Agents*. AGENTS '01. Montreal, Quebec, Canada: ACM, 2001, pp. 385–392. ISBN: 1-58113-326-X. DOI: 10.1145/375735.376343. URL: http://doi.acm.org.libaccess.lib.mcmaster.ca/10.1145/375735.376343.

[248] John E Laird. "Research in human-level AI using computer games". In: *Communications of the ACM* 45.1 (2002), pp. 32–35.

[249] John E Laird. "The Soar Cognitive Architecture". In: (2012).

[250] Richard N. Landers et al. "Theory-driven game-based assessment of general cognitive ability: Design theory, measurement, prediction of performance, and test fairness". In: *Journal of Applied Psychology* 107.10 (Oct. 2022), pp. 1655–1677.

[251] Rogier Landman, Henk Spekreijse, and Victor AF Lamme. "Large capacity storage of integrated objects before change blindness". In: *Vision research* 43.2 (2003), pp. 149–164.

[252] Scott A Langenecker et al. "A task to manipulate attentional load, set-shifting, and inhibitory control: Convergent validity and test–retest reliability of the Parametric Go/No-Go Test". In: *Journal of Clinical and Experimental Neuropsychology* 29.8 (2007), pp. 842–853.

[253] Janet D Larsen, Alan Baddeley, and Jackie Andrade. "Phonological similarity and the irrelevant speech effect: Implications for models of short-term verbal memory". In: *Memory* 8.3 (2000), pp. 145–157.

[254] Nilli Lavie. "Selective attention and cognitive control: Dissociating attentional functions through different types of load". In: *Attention and performance XVIII* (2000), pp. 175–194.

[255] Effie L.-C. Law, Florian Brühlmann, and Elisa D. Mekler. "Systematic Review and Validation of the Game Experience Questionnaire (GEQ) - Implications for Citation and Reporting Practice". In: *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*. CHI PLAY '18. Melbourne, VIC, Australia: Association for Computing Machinery, 2018, 257–270. ISBN: 9781450356244. DOI: 10.1145/3242671.3242683. URL: https://doi.org/10.1145/3242671.3242683.

[256] Robin D Laws. *Robin's Laws of good game mastering*. Steve Jackson Games, 2002.

[257] Nicole Lazzaro. "The Four Fun Keys". In: *Game Usability: Advice from the Experts for Advancing the Player Experience*. Ed. by Katherine Isbister and NoahEditors Schaffer. Morgan Kaufmann, 2008, 317–343.

[258] Cyrille Le Runigo, Nicolas Benguigui, and Benoit G Bardy. "Perception–action coupling and expertise in interceptive actions". In: *Human Movement Science* 24.3 (2005), pp. 429–445.

[259] R A. Leark et al. *Test of Variables of Attention: Clinical Manual*. Los Alamitos, California, USA.: The TOVA Company, 2007.

[260] Christian Lebiere and Robert L West. "A dynamic ACT-R model of simple games". In: *Proceedings of the twenty-first annual conference of the cognitive science society*. Psychology Press. 2020, pp. 296–301.

[261] Marc LeBlanc. *8 Kinds of Fun*. URL: http://algorithmancy.8kindsoffun.com/.

[262] Tyler Lesh et al. "A Multimodal Analysis of Antipsychotic Effects on Brain Structure and Function in First-Episode Schizophrenia". In: *JAMA psychiatry* 72 (Jan. 2015). DOI: 10.1001/jamapsychiatry.2014.2178.

[264] DA Levy, CEL Stark, and LR Squire. "Intact conceptual priming in the absence of declarative memory". In: *Psychological Science* 15.10 (2004), pp. 680–686.

[265] Xinyu Li et al. "To create DDA by the Approach of ANN from UCT-Created Data". In: *2010 international Conference on computer application and system modeling (ICCASM 2010)*. Vol. 8. IEEE. 2010, pp. V8–475.

[266] Ulman Lindenberger, Michael Marsiske, and Paul B Baltes. "Memorizing while walking: increase in dual-task costs from young adulthood to old age." In: *Psychology and aging* 15.3 (2000), p. 417.

[268] Xiaodi Liu et al. "The evaluation of the cognitive and language abilities of autistic children with interactive game technology based on the PEP-3 scale". In: *Education and Information Technologies* (May 2022). ISSN: 1360-2357. DOI: 10.1007/s10639-022-11114-4.

[269] Robert H Logie et al. "Brain activation and the phonological loop: The impact of rehearsal". In: *Brain and Cognition* 53.2 (2003), pp. 293–296.

[270] I Scott MacKenzie. "Fitts' law as a research and design tool in human-computer interaction". In: *Human-computer interaction* 7.1 (1992), pp. 91–139.

[271] Sarah E MacPherson et al. "Specific AD impairment in concurrent performance of two memory tasks". In: *Cortex* 43.7 (2007), pp. 858–865.

[272] Steve Majerus et al. "Exploring the relationship between new word learning and short-term memory for serial order recall, item recall, and item recognition". In: *European Journal of Cognitive Psychology* 18.6 (2006), pp. 848–873.

[273] BR Malcolm et al. "The aging brain shows less flexible reallocation of cognitive resources during dual-task walking: A mobile brain/body imaging (MoBI) study." In: *NeuroImage* 117 (2015), p. 230.

[274] Thomas W Malone. "Toward a theory of intrinsically motivating instruction". In: *Cognitive science* 5.4 (1981), pp. 333–369.

[275] Thomas W. Malone. "What Makes Things Fun to Learn? Heuristics for Designing Instructional Computer Games". In: *Proceedings of the 3rd ACM SIGSMALL Symposium and the First SIGPC Symposium on Small Systems*. SIGSMALL '80. Palo Alto, California, USA: Association for Computing Machinery, 1980, 162–169. ISBN: 0897910249. DOI: 10.1145/800088.802839. URL: https://doi.org/10.1145/800088.802839.

[276] Jess Marcotte. "Queering control (lers) through reflective game design practices". In: *Game Studies* 18.3 (2018), pp. 1–16.

[277] Andrzej Marczewski. "Even Ninja Monkeys like to play". In: *London: Blurb Inc* 1.1 (2015), p. 28.

[278] David Marr and Lucia Vaina. "Representation and recognition of the movements of shapes". In: *Proceedings of the Royal Society of London B: Biological Sciences* 214.1197 (1982), pp. 501–524.

[279] David Marr and A Vision. "A computational investigation into the human representation and processing of visual information". In: *WH San Francisco: Freeman and Company* 1.2 (1982).

[280] Gede Indra Raditya Martha et al. "An Empirical Analysis of Ergonomic Gaming Peripherals Improving Gaming Performance". In: *Journal of Games, Game Art, and Gamification* 7.1 (2022), pp. 15–21.

[284] A. Mavridis and T. Tsiatsos. "Game-based assessment: investigating the impact on test anxiety and exam performance." In: *Journal of Computer Assisted Learning* 33.2 (Apr. 2017), pp. 137–150. DOI: https://doi.org/10.1111/jcal.12170. URL: https://doi.org/10.1111/jcal.12170.

[285] Ashleigh M Maxcey-Richard and Andrew Hollingworth. "The strategic retention of task-relevant objects in visual working memory." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 39.3 (2013), p. 760.

[286] Frans Mäyrä. *An Introduction to Games Studies: Games in Culture*. Sage, 2008.

[287] John McCarthy and Peter Wright. "Technology as experience". In: *interactions* 11.5 (2004), pp. 42–43.

[288] Robert R McCrae and Paul T Costa. "Empirical and theoretical status of the five-factor model of personality traits". In: *The SAGE handbook of personality theory and assessment* 1 (2008), pp. 273–294.

[289] Rory McGloin et al. "Modeling outcomes of violent video game play: Applying mental models and model matching to explain the relationship between user differences, game characteristics, enjoyment, and aggressive intentions". In: *Computers in Human Behavior* 62 (2016), pp. 442–451.

[290] Elinor McKone. "Isolating the special component of face recognition: peripheral identification and a Mooney face." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 30.1 (2004), p. 181.

[291] Elinor McKone, Nancy Kanwisher, and Bradley C Duchaine. "Can generic expertise explain special processing for faces?" In: *Trends in cognitive sciences* 11.1 (2007), pp. 8–15.

[293] Nicole McMahon, Peta Wyeth, and Daniel Johnson. "From Challenges to Activities: Categories of Play in Videogames". In: *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play*. CHI PLAY '15. London, United Kingdom: ACM, 2015, pp. 637–642. ISBN: 978-1-4503-3466-2. DOI: 10.1145/2793107.2810333. URL: http://doi.acm.org/10.1145/2793107.2810333.

[294] Nicole McMahon, Peta Wyeth, and Daniel Johnson. "Personality and player types in Fallout New Vegas". In: *Proceedings of the 4th International Conference on Fun and Games*. FnG '12. Toulouse, France: Association for Computing Machinery, 2012, 113–116. ISBN: 9781450315708. DOI: 10.1145/2367616.2367632. URL: https://doi.org/10.1145/2367616.2367632.

[295] Patricia A McMullen et al. "Apperceptive agnosia and face recognition". In: *Neurocase* 6.5 (2000), pp. 403–414.

[296] Siutila Miia and Havaste Ellinoora. "&ldquoA pure meritocracy blind to identity&rdquo: Exploring the Online Responses to All-Female Teams in Reddit". In: *DiGRA &#3918 - Proceedings of the 2018 DiGRA International Conference: The Game is the Message*. DiGRA, July 2018. URL: http://www.digra.org/wp-content/uploads/digital-library/DIGRA_2018_paper_135.pdf.

[297] Koji Mikami, Kunio Kondo, et al. "Adaptable Game Experience Based on Player's Performance and EEG". In: *2017 Nicograph International (NicoInt)*. IEEE. 2017, pp. 1–8.

[298] Charles Miller et al. "The effect of 5 mechanical gaming keyboard key switch profiles on typing and gaming muscle activity, performance and preferences". In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Vol. 62. 1. SAGE Publications Sage CA: Los Angeles, CA. 2018, pp. 1552–1556.

[299] George A Miller. "The magical number seven, plus or minus two: Some limits on our capacity for processing information." In: *Psychological review* 63.2 (1956), p. 81.

[300] A David Milner and Melvyn A Goodale. "Two visual systems re-viewed". In: *Neuropsychologia* 46.3 (2008), pp. 774–785.

[301] AD Milner and MA Goodale. *Oxford psychology series, No. 27. The visual brain in action*. 1995.

[302] AD Milner et al. "Perception and action in 'visual form agnosia'". In: *Brain* 114.1 (1991), pp. 405–428.

[303] David Milner and Mel Goodale. *The visual brain in action*. Oxford University Press, 2006.

[304] Akira Miyake et al. "The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis". In: *Cognitive psychology* 41.1 (2000), pp. 49–100.

[305] Neville Moray. "Attention in dichotic listening: Affective cues and the influence of instructions". In: *Quarterly journal of experimental psychology* 11.1 (1959), pp. 56–60.

[306] Morris Moscovitch, Gordon Winocur, and Marlene Behrmann. "What is special about face recognition? Nineteen experiments on a person with visual object agnosia and dyslexia but normal face recognition". In: *Journal of cognitive neuroscience* 9.5 (1997), pp. 555–604.

[307] Morris Moscovitch et al. "The cognitive neuroscience of remote episodic, semantic and spatial memory". In: *Current opinion in neurobiology* 16.2 (2006), pp. 179–190.

[308] FM Mottaghy. "Interfering with working memory in humans". In: *Neuroscience* 139.1 (2006), pp. 85–90.

[311] Shane T Mueller and Brian J Piper. "The psychology experiment building language (PEBL) and PEBL test battery". In: *Journal of neuroscience methods* 222 (2014), pp. 250–259.

[312] Shane T Mueller et al. "Theoretical implications of articulatory duration, phonological similarity, and phonological complexity in verbal working memory." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 29.6 (2003), p. 1353.

[314] David Thomas Murphy. "Hybrid moments: Using ludonarrative dissonance for political critique". In: *Loading...* 10.15 (2016).

[315] Lennart Nacke and Anders Drachen. "Towards a framework of player experience research". In: *Proceedings of the second international workshop on evaluating player experience in games at FDG*. Vol. 11. 2011.

[316] Lennart Nacke and Craig A. Lindley. "Flow and Immersion in First-Person Shooters: Measuring the Player's Gameplay Experience". In: *Proceedings of the 2008 Conference on Future Play: Research, Play, Share*. Future Play '08. Toronto, Ontario, Canada: Association for Computing Machinery, 2008, 81–88. ISBN: 9781605582184. DOI: 10.1145/1496984.1496998. URL: https://doi.org/10.1145/1496984.1496998.

[317] Lennart Nacke et al. "Playability and player experience research". In: *Proceedings of digra 2009: Breaking new ground: Innovation in games, play, practice and theory*. DiGRA. 2009.

[318] Lennart E Nacke, Chris Bateman, and Regan L Mandryk. "BrainHex: A neurobiological gamer typology survey". In: *Entertainment Computing* 5.1 (2014), pp. 55–62.

[319] Lennart E Nacke, Anders Drachen, and Stefan Göbel. "Methods for evaluating gameplay experience in a serious gaming context". In: *International Journal of Computer Science in Sport* 9.2 (2010), pp. 1–12.

[320] Daniel Natapov, Steven J. Castellucci, and I. Scott MacKenzie. "ISO 9241-9 Evaluation of Video Game Controllers". In: *Proceedings of Graphics Interface 2009*. GI '09. Kelowna, British Columbia, Canada: Canadian Information Processing Society, 2009, pp. 223–230. ISBN: 978-1-56881-470-4. URL: http://dl.acm.org/citation.cfm?id=1555880.1555930.

[321] Daniel Natapov and I. Scott MacKenzie. "The Trackball Controller: Improving the Analog Stick". In: *Proceedings of the International Academic Conference on the Future of Game Design and Technology*. Futureplay '10. Vancouver, British Columbia, Canada: ACM, 2010, pp. 175–182. ISBN: 978-1-4503-0235-7. DOI: 10.1145/1920778.1920803. URL: http://doi.acm.org/10.1145/1920778.1920803.

[325] Allen Newell, Herbert Alexander Simon, et al. *Human problem solving*. Vol. 104. 9. Prentice-hall Englewood Cliffs, NJ, 1972.

[326] Simon Niedenthal. "What we talk about when we talk about game aesthetics". In: *Digital Games Research Association (DiGRA), London, UK (2009)*. DiGRA Online Library. 2009.

[349] Amelie Nolte et al. "Implementing Ability-Based Design: A Systematic Approach to Conceptual User Modeling". In: *ACM Trans. Access. Comput.* 15.4 (Oct. 2022). ISSN: 1936-7228. DOI: 10.1145/3551646. URL: https://doi.org/10.1145/3551646.

[350] Don Norman. *The design of everyday things: Revised and expanded edition*. Basic books, 2013.

[351] Donald A Norman and Daniel G Bobrow. "On data-limited and resource-limited processes". In: *Cognitive psychology* 7.1 (1975), pp. 44–64.

[352] Joel Norman. "Two visual systems and two theories of perception: An attempt to reconcile the constructivist and ecological approaches". In: *Behavioral and brain sciences* 25.1 (2002), pp. 73–96.

[353] Heather L O'Brien and Elaine G Toms. "The development and evaluation of a survey to measure user engagement". In: *Journal of the American Society for Information Science and Technology* 61.1 (2010), pp. 50–69.

[354] Heather L. O'Brien and Elaine G. Toms. "What is user engagement? A conceptual framework for defining user engagement with technology". In: *Journal of the American Society for Information Science and Technology* 59.6 (2008), pp. 938–955. DOI: https://doi.org/10.1002/asi.20801. eprint: https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/asi.20801. URL: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.20801.

[356] Jacob Kaae Olesen, Georgios N Yannakakis, and John Hallam. "Real-time challenge balance in an RTS game using rtNEAT". In: *2008 IEEE Symposium On Computational Intelligence and Games*. IEEE. 2008, pp. 87–94.

[357] Judith Reitman Olson and Gary M Olson. "The growth of cognitive modeling in human-computer interaction since GOMS". In: *Readings in Human–Computer Interaction*. Elsevier, 1995, pp. 603–625.

[358] Zoë O'Shea and Jonathan Freeman. "Game Design Frameworks: Where Do We Start?" In: *Proceedings of the 14th International Conference on the Foundations of Digital Games*. FDG '19. San Luis Obispo, California, USA: Association for Computing Machinery, 2019. ISBN: 9781450372176. DOI: 10.1145/3337722.3337753. URL: https://doi.org/10.1145/3337722.3337753.

[359] Benjamin Paaßen, Thekla Morgenroth, and Michelle Stratemeyer. "What is a true gamer? The male gamer stereotype and the marginalization of women in video game culture". In: *Sex Roles* 76 (2017), pp. 421–435.

[360] Selma Papegaaij et al. "Aging causes a reorganization of cortical and spinal control of posture". In: *Frontiers in Aging Neuroscience* 6 (2014), p. 28. ISSN: 1663-4365. DOI: 10.3389/fnagi.2014.00028. URL: https://www.frontiersin.org/article/10.3389/fnagi.2014.00028.

[361] Park Associates. *Survey reveals U.S. gamers market is diversifying*. Parks Associates Press Release. Aug. 29. URL: http://www.parksassociates.com/press/press_releases/2006/gaming_pr4.html.

[362] Douglas A Parry et al. "A systematic review and meta-analysis of discrepancies between logged and self-reported digital media use". In: *Nature Human Behaviour* 5.11 (2021), pp. 1535–1547.

[364] Karalyn Patterson, Peter J Nestor, and Timothy T Rogers. "Where do you know what you know? The representation of semantic knowledge in the human brain". In: *Nature Reviews Neuroscience* 8.12 (2007), p. 976.

[365] Christopher A Paul. *The toxic meritocracy of video games: Why gaming culture is the worst*. U of Minnesota Press, 2018.

[366] V. Gregory Payne and Larry David Isaacs. *Human motor development: a lifespan approach*. 8th ed. McGraw-Hill, 2012.

[367] Chris Pedersen, Julian Togelius, and Georgios N Yannakakis. "Modeling player experience in super mario bros". In: *2009 IEEE Symposium on Computational Intelligence and Games*. IEEE. 2009, pp. 132–139.

[368] Peter Peduzzi et al. "A simulation study of the number of events per variable in logistic regression analysis". In: *Journal of Clinical Epidemiology* 49.12 (1996), pp. 1373–1379. ISSN: 0895-4356. DOI: https://doi.org/10.1016/S0895-4356(96)00236-3. URL: https://www.sciencedirect.com/science/article/pii/S0895435696002363.

[369] Federico Peinado and Pablo Gervás. "Transferring game mastering laws to interactive digital storytelling". In: *International Conference on Technologies for Interactive Digital Storytelling and Entertainment*. Springer. 2004, pp. 48–54.

[370] Wilder Penfield and Theodore Rasmussen. "The cerebral cortex of man; a clinical study of localization of function." In: (1950).

[371] Iris-Katharina Penner et al. "The Stroop task: comparison between the original paradigm and computerized versions in children and adults". In: *The Clinical Neuropsychologist* 26.7 (2012), pp. 1142–1153.

[372] M-T Perenin and A Vighetto. "Optic ataxia A specific disruption in visuomotor mechanisms I. Different aspects of the deficit in reaching for objects". In: *Brain* 111.3 (1988), pp. 643–674.

[374] Brian Piper et al. "Evaluation of the validity of the Psychology Experiment Building Language tests of vigilance, auditory memory, and decision making". In: *PeerJ* 4 (2016), e1772.

[377] Susanne Poeller and Cody J. Phillips. "Self-Determination Theory — I Choose You! The Limitations of Viewing Motivation in HCI Research Through the Lens of a Single Theory". In: *Extended Abstracts of the 2022 Annual Symposium on Computer-Human Interaction in Play*. CHI PLAY '22. Bremen, Germany: Association for Computing Machinery, 2022, 261–262. ISBN: 9781450392112. DOI: 10.1145/3505270.3558361. URL: https://doi-org.libaccess.lib.mcmaster.ca/10.1145/3505270.3558361.

[378] K. Poels, Y.A.W. de Kort, and W.A. IJsselsteijn. *D3.3 : Game Experience Questionnaire: development of a self-report measure to assess the psychological impact of digital games*. English. 2007.

[379] Russell A Poldrack and John DE Gabrieli. "Characterizing the neural mechanisms of skill learning and repetition priming: evidence from mirror reading". In: *Brain* 124.1 (2001), pp. 67–82.

[380] Russell A Poldrack et al. "The relationship between skill learning and repetition priming: Experimental and computational analyses." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 25.1 (1999), p. 208.

[381] Teddy Pozo. "Queer games after empathy: Feminism and haptic game design aesthetics from consent to cuteness to the radically soft". In: *Game Studies* 18.3 (2018).

[382] Steven E Prince, Takashi Tsukiura, and Roberto Cabeza. "Distinguishing the neural correlates of episodic memory encoding and semantic memory retrieval". In: *Psychological Science* 18.2 (2007), pp. 144–151.

[383] Dongxiao Qin. "Positionality". In: *The Wiley Blackwell encyclopedia of gender and sexuality studies* (2016), pp. 1–2.

[384] Philip Quinlan, Philip T Quinlan, and Ben Dyson. *Cognitive psychology*. Pearson Education, 2008.

[385] Rishi Rajalingham, Kailyn Schmidt, and James J DiCarlo. "Comparison of object recognition behavior in human and monkey". In: *Journal of Neuroscience* 35.35 (2015), pp. 12127–12136.

[386] Paul Ralph and Kafui Monu. *A Working Theory of Game Design - First Person Scholar*. Sept. 2020. URL: http://www.firstpersonscholar.com/a-working-theory-of-game-design/.

[389] Megan E Renna et al. "The use of the mirror tracing persistence task as a measure of distress tolerance in generalized anxiety disorder". In: *Journal of Rational-Emotive & Cognitive-Behavior Therapy* 36.1 (2018), pp. 80–88.

[392] Charles Reynaldo et al. "Using video games to improve capabilities in decision making and cognitive skill: A literature review". In: *Procedia Computer Science* 179 (2021), pp. 211–221.

[393] Giovanni Ribeiro et al. "Game atmosphere: effects of audiovisual thematic cohesion on player experience and psychophysiology". In: *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. 2020, pp. 107–119.

[394] Jennifer J Richler, Jeremy B Wilmer, and Isabel Gauthier. "General object recognition is specific: Evidence from novel and familiar objects". In: *Cognition* 166 (2017), pp. 42–55.

[396] Raquel Robinson et al. "" Let's Get Physiological, Physiological!" A Systematic Review of Affective Gaming". In: *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. 2020, pp. 132–147.

[397] Paul B Rock, Mike G Harris, and Tim Yates. "A test of the tau-dot hypothesis of braking control in the real world." In: *Journal of Experimental Psychology: Human Perception and Performance* 32.6 (2006), p. 1479.

[398] Ryan Rogers, Nicholas David Bowman, and Mary Beth Oliver. "It's not the model that doesn't fit, it's the controller! The role of cognitive skills in understanding the links between natural mapping, performance, and enjoyment of console video games". In: *Computers in Human Behavior* 49 (2015), pp. 588–596.

[399] Maria Francesca Roig-Maimó et al. "Kill Two Stacks with One Stone:" Fitts" Yourself while Doing Rehabilitation". In: *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 2024, pp. 1–7.

[400] Bruno Rossion and Isabel Gauthier. "How does the brain process upright and inverted faces?" In: *Behavioral and cognitive neuroscience reviews* 1.1 (2002), pp. 63–75.

[401] Mary Rudner et al. "Neural representation of binding lexical signs and words in the episodic buffer of working memory". In: *Neuropsychologia* 45.10 (2007), pp. 2258–2276.

[402] Hans Rutrecht et al. "Time speeds up during flow states: A study in virtual reality with the video game thumper". In: *Timing & Time Perception* 9.4 (2021), pp. 353–376.

[403] Richard M Ryan, C Scott Rigby, and Andrew Przybylski. "The motivational pull of video games: A self-determination theory approach". In: *Motivation and emotion* 30.4 (2006), pp. 344–360.

[404] Joseph B Sala, Pia Rämä, and Susan M Courtney. "Functional topography of a distributed neural system for spatial and nonspatial information maintenance in working memory". In: *Neuropsychologia* 41.3 (2003), pp. 341–356.

[405] Katie Salen, Katie Salen Tekinbaş, and Eric Zimmerman. *Rules of play: Game design fundamentals*. MIT press, 2004.

[406] José Luis González Sánchez et al. "Playability as Extension of Quality in Use in Video Games." In: *I-USED*. 2009.

[407] Fabio Sani and John Todman. *Experimental design and statistics for psychology: a first course*. John Wiley & Sons, 2008.

[408] Thomas Sanocki et al. "Are edges sufficient for object recognition". In: *Journal of Experimental Psychology: Human Perception and Performance* 24.1 (1998), p. 340.

[409] Tiago Santos et al. "What's in a Review: Discrepancies Between Expert and Amateur Reviews of Video Games on Metacritic". In: *Proc. ACM Hum.-Comput. Interact.* 3.CSCW (Nov. 2019). DOI: 10.1145/3359242. URL: https://doi.org/10.1145/3359242.

[410] GJ Savelsbergh, HT Whiting, and Reinoud J Bootsma. "Grasping tau." In: *Journal of Experimental Psychology: Human Perception and Performance* 17.2 (1991), p. 315.

[411] GJP Savelsbergh et al. "The visual guidance of catching". In: *Experimental Brain Research* 93.1 (1993), pp. 148–156.

[414] Daniel L Schacter, Gagan S Wig, and W Dale Stevens. "Reductions in cortical activity during priming". In: *Current opinion in neurobiology* 17.2 (2007), pp. 171–176.

[415] Jesse Schell. *The Art of Game Design: A book of lenses*. CRC Press, 2014.

[416] Björn H Schott et al. "Neuroanatomical dissociation of encoding processes related to priming and explicit memory". In: *Journal of Neuroscience* 26.3 (2006), pp. 792–800.

[417] Björn H Schott et al. "Redefining implicit and explicit memory: the functional neuroanatomy of priming, remembering, and control of retrieval". In: *Proceedings of the National Academy of Sciences* 102.4 (2005), pp. 1257–1262.

[424] CV Segundo et al. "Dynamic difficulty adjustment through parameter manipulation for Space Shooter game". In: *Proceedings of SB Games* (2016).

[425] Rachael D Seidler et al. "Motor control and aging: links to age-related brain structural, functional, and biochemical effects". In: *Neuroscience & Biobehavioral Reviews* 34.5 (2010), pp. 721–733.

[426] Kuppuraj Sengottuvel and Prema KS Rao. "An adapted serial reaction time task for sequence learning measurements". In: *Psychological Studies* 58.3 (2013), pp. 276–284.

[427] Steven C Seow. "Information theoretic models of HCI: A comparison of the Hick-Hyman law and Fitts' law". In: *Human-computer interaction* 20.3 (2005), pp. 315–352.

[428] David Servan-Schreiber, Jonathan D Cohen, and Sandra Steingard. "Schizophrenic deficits in the processing of context: A test of a theoretical model". In: *Archives of general psychiatry* 53.12 (1996), pp. 1105–1112.

[429] David R Shaffer and Katherine Kipp. *Developmental psychology: Childhood and adolescence*. 8th ed. Cengage Learning, 2010.

[430] Noor Shaker, Georgios Yannakakis, and Julian Togelius. "Towards automatic personalized content generation for platform games". In: *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*. Vol. 5. 1. 2010.

[431] Adrienne Shaw. "Do you identify as a gamer? Gender, race, sexuality, and gamer identity". In: *new media & society* 14.1 (2012), pp. 28–44. DOI: 10.1177/1461444811410394. eprint: https://doi.org/10.1177/1461444811410394. URL: https://doi.org/10.1177/1461444811410394.

[432] Adrienne Shaw. "On not becoming gamers: Moving beyond the constructed audience". In: (2013).

[433] David K Sherman and Geoffrey L Cohen. "The psychology of self-defense: Self-affirmation theory". In: *Advances in experimental social psychology* 38 (2006), pp. 183–242.

[434] Miguel Sicart. "The ethics of computer games." In: (2009).

[435] J. Richard Simon. "Steadiness, Handedness, and Hand Preference". In: *Perceptual and Motor Skills* 18.1 (1964). PMID: 14116331, pp. 203–206. DOI: 10.2466/pms.1964.18.1.203. eprint: https://doi.org/10.2466/pms.1964.18.1.203. URL: https://doi.org/10.2466/pms.1964.18.1.203.

[437] Mario Luis Small. "How to conduct a mixed methods study: Recent trends in a rapidly growing literature". In: *Annual review of sociology* 37 (2011), pp. 57–86.

[438] Edward E Smith and John Jonides. "Working memory: A view from neuroimaging". In: *Cognitive psychology* 33.1 (1997), pp. 5–42.

[439] Javier Snaider, Ryan McCall, and Stan Franklin. "The LIDA framework as a general tool for AGI". In: *Artificial general intelligence* (2011), pp. 133–142.

[441] S. Soraine and J. Carette. "Mechanical Experience, Competency Profiles, and Jutsu". In: *Journal of Games, Self, & Society* 2.1 (Apr. 2020), pp. 150–207. DOI: 10.1184/R1/12215417.v1. URL: https://kilthub.cmu.edu/articles/Journal_of_Games_Self_Society_Vol_2_No_1/12215417.

[442] Sasha M Soraine. "Ooh What's This Button Do?" MA thesis. McMaster University, Mar. 2019. URL: http://hdl.handle.net/11375/24028.

[443] George Sperling. "The information available in brief visual presentations." In: *Psychological monographs: General and applied* 74.11 (1960), p. 1.

[444] Hugo J Spiers, Eleanor A Maguire, and Neil Burgess. "Hippocampal amnesia". In: *Neurocase* 7.5 (2001), pp. 357–382.

[445] Pieter Spronck, Ida Sprinkhuizen-Kuyper, and Eric Postma. "Online adaptation of game opponent AI with dynamic scripting". In: *International Journal of Intelligent Games and Simulation* 3.1 (2004), pp. 45–53.

[448] Claude M Steele. "The psychology of self-affirmation: Sustaining the integrity of the self". In: *Advances in experimental social psychology*. Vol. 21. Elsevier, 1988, pp. 261–302.

[449] Adi Stein et al. "EEG-triggered dynamic difficulty adjustment for multiplayer games". In: *Entertainment computing* 25 (2018), pp. 14–25.

[450] Randy Stein and Alexander B Swan. "Evaluating the validity of Myers-Briggs Type Indicator theory: A teaching tool and window into intuitive psychology". In: *Social and Personality Psychology Compass* 13.2 (2019), e12434.

[451] Bart Stewart. *Personality And Play Styles: A Unified Model*. Sept. 2011. URL: https://www.gamasutra.com/view/feature/6474/per-sonality_and_play_styles_a_.php.

[452] Jeffrey A Stone. "Self-identification as a "gamer" among college students: Influencing factors and perceived characteristics". In: *New media & society* 21.11-12 (2019), pp. 2607–2627.

[453] Tilo Strobach, Peter A Frensch, and Torsten Schubert. "Video game practice optimizes executive control skills in dual-task and task switching situations". In: *Acta psychologica* 140.1 (2012), pp. 13–24.

[454] Donald T Stuss and Michael P Alexander. "Is there a dysexecutive syndrome?" In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 362.1481 (2007), pp. 901–915.

[455] Ron Sun. "The CLARION cognitive architecture: Extending cognitive modeling to social simulation". In: *Cognition and multi-agent interaction* (2006), pp. 79–99.

[456] Ron Sun. "The importance of cognitive architectures: an analysis based on CLARION". In: *Journal of Experimental & Theoretical Artificial Intelligence* 19.2 (June 2007), 159–193. DOI: 10.1080/09528130701191560.

[457] Penelope Sweetser and Peta Wyeth. "GameFlow: A Model for Evaluating Player Enjoyment in Games". In: *Comput. Entertain.* 3.3 (July 2005), pp. 3–3. ISSN: 1544-3574. DOI: 10.1145/1077246.1077253. URL: http://doi.acm.org/10.1145/1077246.1077253.

[458] Penelope Sweetser et al. "GameFlow in Different Game Genres and Platforms". In: *Comput. Entertain.* 15.3 (Apr. 2017). DOI: 10.1145/3034780. URL: https://doi.org/10.1145/3034780.

[459] Penny Sweetser, Daniel Johnson, and Peta Wyeth. "Revisiting the GameFlow model with detailed heuristics". In: *Journal of Creative Technologies* 2012.3 (2012), pp. 1–16.

[460] John Sweller. "Cognitive load during problem solving: Effects on learning". In: *Cognitive science* 12.2 (1988), pp. 257–285.

[461] José Luis González Sánchez et al. "Playability: analysing user experience in video games". In: *Behaviour & Information Technology* 31.10 (2012), pp. 1033–1054. DOI: 10.1080/0144929X.2012.710648. eprint: https://doi.org/10.1080/0144929X.2012.710648. URL: https://doi.org/10.1080/0144929X.2012.710648.

[462] Ron Tamborini. "The experience of telepresence in violent video games". In: *86th annual convention of the National Communication Association, Seattle, WA*. 2000.

[463] Ron Tamborini and Paul Skalski. "The role of presence in the experience of electronic games". In: *Playing video games: Motives, responses, and consequences* 1 (2006), pp. 225–240.

[464] C. H. Tan, K. C. Tan, and A. Tay. "Dynamic Game Difficulty Scaling Using Adaptive Behavior-Based AI". In: *IEEE Transactions on Computational Intelligence and AI in Games* 3.4 (2011), pp. 289–301. DOI: 10.1109/TCIAIG.2011.2158434.

[465] James W Tanaka and Martha J Farah. "The holistic representation of faces". In: *Perception of faces, objects, and scenes: Analytic and holistic processes* (2003), pp. 53–74.

[467] Thomas Terkildsen, Lene Engelst, and Mathias Clasen. "Work Hard, Scare Hard? An Investigation of How Mental Workload Impacts Jump Scare Intensity". In: *Proc. ACM Hum.-Comput. Interact.* 7.CHI PLAY (Oct. 2023). DOI: 10.1145/3611021. URL: https://doi.org/10.1145/3611021.

[468] Mateo Terrasa-Torres et al. "Difficulty as Aesthetic: An Investigation of the Expressiveness of Challenge in Digital Games". In: *Acta Ludologica* 4.1 (2021), pp. 94–111.

[469] Gareth Terry et al. "Thematic analysis". In: *The SAGE handbook of qualitative research in psychology* 2.17-37 (2017), p. 25.

[470] Hugh R Terry, Samuel G Charlton, and John A Perrone. "The role of looming and attention capture in drivers' braking responses". In: *Accident Analysis & Prevention* 40.4 (2008), pp. 1375–1382.

[471] Jan-Noël Thon. "Playing with Fear: The Aesthetics of Horror in Recent Indie Games". In: *Eludamos: Journal for Computer Game Culture* 10.1 (2019), pp. 197–231.

[472] David Thue et al. "Interactive storytelling: A player modelling approach". In: *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*. Vol. 3. 1. 2007, pp. 43–48.

[473] Gustavo F. Tondello et al. "The Gamification User Types Hexad Scale". In: *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play*. CHI PLAY '16. Austin, Texas, USA: Association for Computing Machinery, 2016, 229–243. ISBN: 9781450344562. DOI: 10.1145/2967934.2968082. URL: https://doi.org/10.1145/2967934.2968082.

[474] Gustavo F. Tondello et al. "Towards a trait model of video game preferences". In: *International Journal of Human–Computer Interaction* 34.8 (2018), pp. 732–748.

[475] Gustavo F Tondello et al. ""I don't fit into a single type": A Trait Model and Scale of Game Playing Preferences". In: *IFIP Conference on Human-Computer Interaction*. Springer. 2019, pp. 375–395.

[476] Anne M Treisman. "Contextual cues in selective listening". In: *Quarterly Journal of Experimental Psychology* 12.4 (1960), pp. 242–248.

[477] Anne M Treisman. "Verbal cues, language, and meaning in selective attention". In: *The American journal of psychology* 77.2 (1964), pp. 206–219.

[478] Jochen Triesch et al. "What you see is what you need". In: *Journal of vision* 3.1 (2003), pp. 9–9.

[479] Luigi Trojano and Dario Grossi. "Phonological and lexical coding in verbal short-term-memory and learning". In: *Brain and language* 51.2 (1995), pp. 336–354.

[480] Fan-Chen Tseng. "Segmenting online gamers by motivation". In: *Expert Systems with Applications* 38.6 (2011), pp. 7693–7697.

[481] Endel Tulving. "Episodic memory: From mind to brain". In: *Annual review of psychology* 53.1 (2002), pp. 1–25.

[482] Endel Tulving and Daniel L Schacter. "Priming and human memory systems". In: *Science* 247.4940 (1990), p. 301.

[483] Endel Tulving et al. "Episodic and semantic memory". In: *Organization of Memory* 1 (1972), pp. 381–403.

[484] J.C Turner et al. *Rediscovering the social group: A self-categorization theory.* Oxford, England: Blackwell, 1987.

[485] April Tyack and Elisa D. Mekler. "Off-Peak: An Examination of Ordinary Player Experience". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI '21. Yokohama, Japan: Association for Computing Machinery, 2021. ISBN: 9781450380966. DOI: 10.1145/3411764.3445230. URL: https://doi.org/10.1145/3411764.3445230.

[486] April Tyack and Elisa D. Mekler. "Self-Determination Theory in HCI Games Research: Current Uses and Open Questions". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20. Honolulu, HI, USA: Association for Computing Machinery, 2020, 1–22. ISBN: 9781450367080. DOI: 10.1145/3313831.3376723. URL: https://doi-org.libaccess.lib.mcmaster.ca/10.1145/3313831.3376723.

[489] Nash Unsworth and Randall W Engle. "Individual differences in working memory capacity and learning: Evidence from the serial reaction time task". In: *Memory & cognition* 33.2 (2005), pp. 213–220.

[490] Giuseppe Vallar and Costanza Papagno. "Neuropsychological impairments of short-term memory." In: (1995).

[494] Michael Van Lent et al. "Intelligent agents in computer games". In: *AAAI/IAAI*. 1999, pp. 929–930.

[495] A. Vandierendonck. "A comparison of methods to combine speed and accuracy measures of performance: A rejoinder on the binning procedure". In: *Behaviour Methods Research* 49 (2 Mar. 2017), pp. 653–673. DOI: 10.3758/s13428-016-0721-5.

[496] Cinnamon VanPutte, Jennifer Regan, and Andrew Russo. *Seeley's Anatomy and Physiology*. 9th ed. McGraw-Hill, 2011.

[497] Faraneh Vargha-Khadem, David G Gadian, and Mortimer Mishkin. "Dissociations in cognitive memory: the syndrome of developmental amnesia". In: *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 356.1413 (2001), pp. 1435–1440.

[498] Faraneh Vargha-Khadem et al. "Differential effects of early hippocampal pathology on episodic and semantic memory". In: *Science* 277.5324 (1997), pp. 376–380.

[505] Karhulahti Veli-Matti. "Puzzle Is Not a Game! Basic Structures of Challenge". In: *Proceedings of the 2013 DiGRA International Conference: DeFragging Game Studies*. DiGRA, Aug. 2014. ISBN: ISSN 2342-9666. URL: http://www.digra.org/wp-content/uploads/digital-library/paper_179.pdf.

[506] Matthew Ventura, Valerie Shute, and Weinan Zhao. "The relationship between video game use and a performance-based measure of persistence". In: *Computers & Education* 60.1 (2013), pp. 52–58.

[507] M Verfaellie, P Koseff, and MP Alexander. "Acquisition of novel semantic information in amnesia: effects of lesion location". In: *Neuropsychologia* 38.4 (2000), pp. 484–492.

[508] Lotte Vermeulen, Sofie Van Bauwel, and Jan Van Looy. "Tracing female gamer identity. An empirical study into gender and stereotype threat perceptions". In: *Computers in Human Behavior* 71 (2017), pp. 90–98.

[509] Lotte Vermeulen and Jan Van Looy. ""I play so I am?" A gender study into stereotype perception and genre choice of digital game players". In: *Journal of Broadcasting & Electronic Media* 60.2 (2016), pp. 286–304.

[510] Lotte Vermeulen et al. "Playing under threat. Examining stereotype threat in female game players". In: *Computers in Human Behavior* 57 (2016), pp. 377–387.

[511] Edward K Vogel, Geoffrey F Woodman, and Steven J Luck. "Storage of features, conjunctions, and objects in visual working memory." In: *Journal of experimental psychology: human perception and performance* 27.1 (2001), p. 92.

[512] Wolfgang Walk, Daniel Görlich, and Mark Barrett. "Design, Dynamics, Experience (DDE): an advancement of the MDA framework for game design". In: *Game Dynamics*. Springer, 2017, pp. 27–45.

[513] Russell T Warne. *Statistics for the social sciences: a general linear model approach*. 1st ed. Cambridge, United Kingdom: Cambridge University Press, 2017.

[514] Elizabeth K Warrington. "Visual object and space perception battery". In: *(No Title)* (1991).

[515] Wayne Weiten. *Psychology Themes and variations*. Cengage Learning, 2007.

[516] Mark A Wheeler, Donald T Stuss, and Endel Tulving. "Toward a theory of episodic memory: the frontal lobes and autonoetic consciousness." In: *Psychological bulletin* 121.3 (1997), p. 331.

[517] Christopher D. Wickens. "Multiple Resources and Mental Workload". In: *Human Factors* 50.3 (2008). PMID: 18689052, pp. 449–455. DOI: 10.1518/001872008X288394. eprint: https://doi.org/10.1518/001872008X288394. URL: https://doi.org/10.1518/001872008X288394.

[518] Christopher D Wickens et al. *An Introduction to Human Factors Engineering*. 2nd ed. Harlow, Essex, England: Pearson Education, 2014.

[519] Eric N. Wiebe et al. "Measuring engagement in video game-based environments: Investigation of the User Engagement Scale". In: *Computers in Human Behavior* 32 (2014), pp. 123–132. ISSN: 0747-5632. DOI: https://doi.org/10.1016/j.chb.2013.12.001. URL: https://www.sciencedirect.com/science/article/pii/S0747563213004433.

[520] Gagan S Wig et al. "Reductions in neural activity underlie behavioral components of repetition priming". In: *Nature neuroscience* 8.9 (2005), p. 1228.

[521] Richard Wilkie and John Wann. "Controlling steering and judging heading: retinal flow, visual direction, and extraretinal information." In: *Journal of Experimental Psychology: Human Perception and Performance* 29.2 (2003), p. 363.

[522] Richard M Wilkie, John P Wann, and Robert S Allison. "Active gaze, visual look-ahead, and locomotor control". In: *Journal of experimental psychology: Human perception and performance* 34.5 (2008), pp. 1150–1164.

[523] Dmitri Williams, Nick Yee, and Scott E Caplan. "Who plays, how much, and why? Debunking the stereotypical gamer profile". In: *Journal of computer-mediated communication* 13.4 (2008), pp. 993–1018.

[524] Russell B Williams. "Conceptual models and mental models in operation: Frustration, performance and flow with two different video game controllers". In: *Entertainment Computing* 28 (2018), pp. 2–10.

[525] Brian M Winn. "The design, play, and experience framework". In: *Handbook of research on effective electronic gaming in education*. IGI Global, 2009, pp. 1010–1024.

[526] Hanna Wirman. ""I am not a fan, I just play a lot"–If Power Gamers Aren't Fans, Who Are?" In: *Proceedings of DiGRA 2007 Conference: Situated Play*. 2007.

[527] W Wirth et al. "Presence as a process: Towards a unified theoretical model of formation of spatial presence experiences". In: *Unpublished manuscript* (2003).

[529] D.J. Woltz and C.A. Was. "Availability of related long-term memory during and after attention focus in working memory". In: *Memory & Cognition* 34 (Apr. 2006), pp. 668–684. DOI: 10.3758/BF03193587.

[530] Alan C-N Wong, Thomas J Palmeri, and Isabel Gauthier. "Conditions for facelike expertise with objects: Becoming a Ziggerin expert—but which type?" In: *Psychological Science* 20.9 (2009), pp. 1108–1117.

[532] Peter Wright, John McCarthy, and Lisa Meekison. "Making sense of experience". In: *Funology: From usability to enjoyment*. Springer, 2003, pp. 43–53.

[533] Su Xue et al. "Dynamic difficulty adjustment for maximized engagement in digital games". In: *Proceedings of the 26th International Conference on World Wide Web Companion*. 2017, pp. 465–471.

[534] Nick Yee. "Motivations for play in online games". In: *CyberPsychology & behavior* 9.6 (2006), pp. 772–775.

[535] Nick Yee. "The Gamer Motivation Profile: What We Learned From 250,000 Gamers". In: *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play*. CHI PLAY '16. Austin, Texas, USA: Association for Computing Machinery, 2016, p. 2. ISBN: 9781450344562. DOI: 10.1145/2967934.2967937. URL: https://doi.org/10.1145/2967934.2967937.

[536] Nick Yee, Nicolas Ducheneaut, and Les Nelson. "Online gaming motivations scale: development and validation". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '12. Austin, Texas, USA: Association for Computing Machinery, 2012, 2803–2806. ISBN: 9781450310154. DOI: 10.1145/2207676.2208681. URL: https://doi.org/10.1145/2207676.2208681.

[537] Emre H Yilmaz and William H Warren. "Visual control of braking: A test of the t hypothesis." In: *Journal of Experimental Psychology: Human Perception and Performance* 21.5 (1995), p. 996.

[538] Galit Yovel and Brad Duchaine. "Specialized face perception mechanisms extract both part and spacing information: Evidence from developmental prosopagnosia". In: *Journal of Cognitive Neuroscience* 18.4 (2006), pp. 580–593.

[539] Chong Ho Yu. "Test–Retest Reliability". In: *Encyclopedia of Social Measurement*. Ed. by Kimberly Kempf-Leonard. New York: Elsevier, 2005, pp. 777–784. ISBN: 978-0-12-369398-3. DOI: https://doi.org/10.1016/B0-12-369398-5/00094-3. URL: https://www.sciencedirect.com/science/article/pii/B0123693985000943.

[540] José P Zagal. "A framework for games literacy and understanding games". In: *Proceedings of the 2008 Conference on Future Play: Research, Play, Share*. 2008, pp. 33–40.

[541] Myrka Zago et al. "Internal models and prediction of visual gravitational motion". In: *Vision research* 48.14 (2008), pp. 1532–1538.

[542] Eric Zimmerman. "Gaming Literacy: Game Design as a Model for Literacy in the Twenty-First Century". In: *The Video Game Theory Reader 2*. Ed. by Bernard Perron. Ed. by Mark JP Wolf. Ed. by Thomas H Apperley. New York, USA: Routledge New York, 2009. Chap. 1, pp. 23–31.

[543] Mohammad Zohaib and Hideyuki Nakanishi. "Dynamic Difficulty Adjustment (DDA) in Computer Games: A Review". In: *Adv. in Hum.-Comp. Int.* 2018 (Jan. 2018). ISSN: 1687-5893. DOI: 10.1155/2018/5681652. URL: https://doi.org/10.1155/2018/5681652.

Journalistic and Community References

[15] Entertainment Software Association. *2023 Essential Facts About the U.S. Video Game Industry*. 2023. URL: https://www.theesa.com/2023-essential-facts/.

[16] Entertainment Software Association. *Real Canadian Gamer Essential Facts 2020*. 2020. URL: https://theesa.ca/wp-content/uploads/2022/10/RCGEF_en.pdf.

[292] Rus McLaughlin. *Heavy rain and redeeming the quick-time event*. Jan. 7. URL: https://venturebeat.com/community/2011/01/07/heavy-rain-and-redeeming-the-quick-time-event/.

[499] Various Reddit Users. *Avoiding adaptive difficulty*. Apr. 2020. URL: https://www.reddit.com/r/residentevil/comments/g6udn2/avoiding_adaptive_difficulty/.

[500] Various Reddit Users. *Dynamic Game Difficulty Adjustment*. Jan. 2012. URL: https://www.reddit.com/r/truegaming/comments/oz14q/dynamic_game_difficulty_adjustment/.

[501] Various Reddit Users. "r/Keyboard: Is mechanical keyboard is necessary for gaming or it is just for aesthetic?" [Online]. Aug. 31. URL: https://www.reddit.com/r/Keyboard/comments/16679av/is_mechanical_keyboard_is_necessary_for_gaming_or/.

[502] Various Reddit Users. *What do you think about auto-dynamic-difficulty?* Apr. 2018. URL: https://www.reddit.com/r/truegaming/comments/89alfu/what_do_you_think_about_autodynamicdifficulty/.

[503] Various Steam Community Users. *Disable Adaptive Difficulty *Updated to work w/ Ghost Survivors Update* :: Resident Evil 2 General Discussions*. Feb. 2019. URL: https://steamcommunity.com/app/883710/discussions/0/1779388024849763978/?l=english.

[531] Kevin Wong. *In Defense of Button Mashing*. Oct. 23. URL: https://kotaku.com/in-defense-of-button-mashing-1648604756.

Video Games

[1] 3909 LLC. Papers, Please. Game [Windows]. 3909 LLC, USA. USA, Aug. 2013.

[9] Alexey Pajitnov. Tetris. Game [Arcade]. SEGA, Tokyo, Japan. Tokyo, Japan, Dec. 1988.

[10] AmberDrop. Try to Fall Asleep. Game [PS4]. AmberDrop, Latvia. Latvia, June 2013.

[17] Atlus. Trauma Center: New Blood. Game [Wii]. Atlus USA, California, USA. California, USA, Nov. 2007.

[18] Atlus. Trauma Center: Second Opinion. Game [Wii]. Atlus USA, California, USA. California, USA, Nov. 2006.

[19] Atlus. Trauma Center: Trauma Team. Game [Wii]. Atlus USA, California, USA. California, USA, May 2010.

[28] Bandai Namco Studios. New Pokemon Snap! Game [Switch]. Nintendo, Kyoto, Japan. Kyoto, Japan, Apr. 2021.

[38] Ben Esposito. Donut County. Game [PS4]. Annapurna Interactive, California, USA. California, USA, Aug. 2018.

[69] Bungie. Halo 3. Game [Xbox 360]. Microsoft Game Studios, Washington, USA. Washington, USA, Sept. 2007.

[75] Campo Santo. Firewatch. Game [Windows]. Panic, Oregon, USA. Oregon, USA, Feb. 2016.

[76] Capcom Production. Resident Evil 4. Game [Gamecube]. Capcom, Osaka, Japan. Osaka, Japan, Jan. 2005.

[77] Capcom Production. Resident Evil 7: Biohazard. Game [PS4]. Capcom, Osaka, Japan. Osaka, Japan, Jan. 2017.

[90] Clover Studio. Okami! Game [Wii]. Capcom, Osaka, Japan. Osaka, Japan, Apr. 2008.

[93] ConcernedApe. Stardew Valley. Game [Windows]. ConcernedApe, California, USA. California, USA, Feb. 2016.

[95] Cooking Mama Ltd. Cooking Mama. Game [Nintendo DS]. Majesco Entertainment, New Jersey, USA. New Jersey, USA, Sept. 2006.

[96] Cooking Mama Ltd. Cooking Mama: Cook Off. Game [Wii]. Majesco Entertainment, New Jersey, USA. New Jersey, USA, Mar. 2007.

[102] Crystal Dynamics. Tomb Raider: Underworld. Game [Xbox 360]. Eidos Interactive, London, UK. London, UK, Nov. 2008.

[128] Extremely OK Games. Celeste. Game [Windows]. Extremely OK Games, British Columbia, Canada. British Columbia, Canada, Jan. 2018.

[131] Facebook Gaming Team. Asteroids Attack. Game [Android]. Facebook, Massachussetts, USA. Massachussetts, USA, Aug. 2018.

[150] Game Freak. Pokemon X. Game [3DS]. Nintendo, Kyoto, Japan. Kyoto, Japan, Oct. 2013.

[151] Game Freak. Pokemon Y. Game [3DS]. Nintendo, Kyoto, Japan. Kyoto, Japan, Oct. 2013.

[155] Ghost Town Games. Overcooked! Game [Switch]. Team17, England. Wakefield, England, July 2017.

[161] Good Science Studios. Kinect Adventures! Game [Xbox 360]. Microsoft Game Studios, Washington, USA. Washington, USA, Nov. 2010.

[175] HAL Laboratory. Kirby: Canvas Curse. Game [Nintendo DS]. Nintendo, Kyoto, Japan. Kyoto, Japan, June 2005.

[176] HAL Laboratory, Inc. and Pax Softnica. Pokemon Snap! Game [N64]. Nintendo, Kyoto, Japan. Kyoto, Japan, June 1999.

[177] Halfbrick Studios. Fruit Ninja. Game [Anrdoid]. Halfbrick Studios, Brisbane, Australia. Brisbane, Australia, Sept. 2010.

[181] Harmonix. Dance Central. Game [Xbox 360]. MTV Games, New York, USA. New York, USA, Nov. 2010.

[193] Hudson Soft. Mario Party. Game [N64]. Nintendo, Kyoto, Japan. Kyoto, Japan, Feb. 1999.

[194] Hudson Soft. Mario Party 2. Game [N64]. Nintendo, Kyoto, Japan. Kyoto, Japan, Jan. 2000.

[195] Hudson Soft. Mario Party 3. Game [N64]. Nintendo, Kyoto, Japan. Kyoto, Japan, May 2001.

[196] Hudson Soft. Mario Party 4. Game [Gamecube]. Nintendo, Kyoto, Japan. Kyoto, Japan, Nov. 2002.

[197] Hudson Soft. Mario Party 5. Game [Gamecube]. Nintendo, Kyoto, Japan. Kyoto, Japan, Nov. 2003.

[198] Hudson Soft. Mario Party 6. Game [Gamecube]. Nintendo, Kyoto, Japan. Kyoto, Japan, Dec. 2004.

[203] id Software. Quake. Game [Windows]. GT Interactive, New York, USA. New York, USA, July 1996.

[208] Intelligent Systems and Nintendo SPD. WarioWare: Touched! Game [DS]. Nintendo, Kyoto, Japan. Kyoto, Japan, Feb. 2005.

[237] Konami. Dance Dance Revolution. Game [Arcade]. Konami, Tokyo, Japan. Tokyo, Japan, Mar. 1999.

[238] Konami. Track & Field. Game [Wii]. Centuri, Florida, USA. Florida, USA, Oct. 1983.

[263] Level-5. Professor Layton and the Diabolical Box. Game [DS]. Nintendo, Kyoto, Japan. Kyoto, Japan, Aug. 2009.

[267] Lionhead Studios. Fable Anniversary. Game [Xbox 360]. Microsoft Studios, Washington, USA. Washington, USA, Feb. 2014.

[322] Nd Cube. Mario Party 9. Game [Wii]. Nintendo, Kyoto, Japan. Kyoto, Japan, Mar. 2012.

[323] Nd Cube. Mario Party Superstars. Game [Switch]. Nintendo, Kyoto, Japan. Kyoto, Japan, Oct. 2021.

[324] Nd Cube. Super Mario Party. Game [Switch]. Nintendo, Kyoto, Japan. Kyoto, Japan, Oct. 2018.

[327] Nintendo EAD. Legend of Zelda: The Wind Waker. Game [Gamecube]. Nintendo, Kyoto, Japan. Kyoto, Japan, Mar. 2003.

[328] Nintendo EAD. Mario Kart 8. Game [WiiU]. Nintendo, Kyoto, Japan. Kyoto, Japan, May 2014.

[329] Nintendo EAD. Super Mario Sunshine. Game [Gamecube]. Nintendo, Kyoto, Japan. Kyoto, Japan, Aug. 2002.

[330] Nintendo EAD. The Legend of Zelda: A Link to the Past. Game [NES]. Nintendo, Kyoto, Japan. Kyoto, Japan, Apr. 1992.

[331] Nintendo EAD. The Legend of Zelda: Majora's Mask. Game [N64]. Nintendo, Kyoto, Japan. Kyoto, Japan, Apr. 2000.

[332] Nintendo EAD. The Legend of Zelda: Skyward Sword. Game [Wii]. Nintendo, Kyoto, Japan. Kyoto, Japan, Nov. 2011.

[333] Nintendo EAD. Wii Sports. Game [Wii]. Nintendo, Kyoto, Japan. Kyoto, Japan, Nov. 2006.

[334] Nintendo EAD and HAL Laboratory, Inc. Pokemon Stadium. Game [N64]. Nintendo, Kyoto, Japan. Kyoto, Japan, Feb. 2000.

[335] Nintendo EAD Group No. 2. Animal Crossing: City Folk. Game [Wii]. Nintendo, Kyoto, Japan. Kyoto, Japan, Nov. 2008.

[336] Nintendo EAD Group No. 2. Wii Sports Resort. Game [Wii]. Nintendo, Kyoto, Japan. Kyoto, Japan, July 2009.

[337] Nintendo EAD Group No. 5. Wii Fit. Game [Wii]. Nintendo, Kyoto, Japan. Kyoto, Japan, May 2008.

[338] Nintendo EAD Group No. 5. Wii Fit Plus. Game [Wii]. Nintendo, Kyoto, Japan. Kyoto, Japan, Oct. 2009.

[339] Nintendo EAD Tokyo. Super Mario Galaxy. Game [Wii]. Nintendo, Kyoto, Japan. Kyoto, Japan, Nov. 2007.

[340] Nintendo Entertaining Planning & Development Division. Ring Fit Adventure. Game [Switch]. Nintendo, Kyoto, Japan. Kyoto, Japan, Oct. 2019.

[341] Nintendo Entertainment Planning & Development Division. Nintendo Switch Sports. Game [Switch]. Nintendo, Kyoto, Japan. Kyoto, Japan, Apr. 2022.

[342] Nintendo EPD. Animal Crossing: New Horizons. Game [Switch]. Nintendo, Kyoto, Japan. Kyoto, Japan, Mar. 2020.

[343] Nintendo EPD. Super Mario Odyssey. Game [Switch]. Nintendo, Kyoto, Japan. Kyoto, Japan, Oct. 2017.

[344] Nintendo EPD. The Legend of Zelda: Breath of the Wild. Game [WiiU]. Nintendo, Kyoto, Japan. Kyoto, Japan, Mar. 2017.

[345] Nintendo EPD. The Legend of Zelda: Breath of the Wild. Game [Switch]. Nintendo, Kyoto, Japan. Kyoto, Japan, Mar. 2017.

[346] Nintendo EPD. The Legend of Zelda: Tears of the Kingdom. Game [Switch]. Nintendo, Kyoto, Japan. Kyoto, Japan, May 2023.

[347] Nintendo R&D2. Super Mario World 2: Yoshi's Island. Game [GBA]. Nintendo, Kyoto, Japan. Kyoto, Japan, Sept. 2002.

[348] Nintendo SPD Group No. 4 and Nd Cube. Wii Party. Game [Wii]. Nintendo, Kyoto, Japan. Kyoto, Japan, Oct. 2010.

[355] Obsidian Entertainment and South Park Digital Studios. South Park: The Stick of Truth. Game [Xbox]. Ubisoft, Rennes, France. Rennes, France, Mar. 2014.

[363] Parsec Productions. Slender: The Eight Pages. Game [PC]. Parsec Productions, New Mexico, USA. New Mexico, USA, June 2012.

[373] Perfect Tap Games. Chicken Scream. Game [Android]. Perfect Tap Games, Dubai, UAE. Dubai, UAE, Apr. 2017.

[375] Platinum Games. Bayonetta. Game [Xbox 360]. Sega, Tokyo, Japan. Tokyo, Japan, Jan. 2010.

[376] Platinum Games. Bayonetta. Game [WiiU]. Nintendo, Kyoto, Japan. Kyoto, Japan, Oct. 2014.

[387] Rare. Kinect Sports. Game [Xbox 360]. Microsoft Game Studios, Washington, USA. Washington, USA, Nov. 2010.

[388] Rare. Sea of Thieves. Game [PC]. Microsoft Studios, Washington, USA. Washington, USA, Mar. 2018.

[390] Retro Studios. Donkey Kong Country: Returns. Game [Wii]. Nintendo, Kyoto, Japan. Kyoto, Japan, Nov. 2010.

[391] Retro Studios. Donkey Kong Country: Tropical Freeze. Game [WiiU]. Nintendo, Kyoto, Japan. Kyoto, Japan, Feb. 2014.

[395] Riot Games. League of Legends. Game [Windows]. Riot Games, California, USA. California, USA, Oct. 2009.

[412] SCE Santa Monica Studio. God of War 2. Game [PS2]. Sony Computer Entertainment, California, USA. California, USA, Mar. 2007.

[413] SCE Santa Monica Studio. God of War 3. Game [PS3]. Sony Computer Entertainment, California, USA. California, USA, Mar. 2010.

[418] Scott Cawthon. Five Nights at Freddy's. Game [Windows]. Scott Cawthon, Texas, USA. Texas, USA, Aug. 2014.

[419] Sega. Mario and Sonic at the Olympic Games Tokyo 2020. Game [Switch]. Sega, Tokyo, Japan. Tokyo, Japan, Nov. 2019.

[420] Sega Sports R&D. Mario and Sonic at the Olympic Games. Game [DS]. Sega, Tokyo, Japan. Tokyo, Japan, Jan. 2008.

[421] Sega Sports R&D and Arzest. Mario and Sonic at the Rio 2016 Olympic Games. Game [3DS]. Nintendo, Kyoto, Japan. Kyoto, Japan, Mar. 2016.

[422] Sega Sports R&D and Racjin. Mario and Sonic at the Winter Olympic Games. Game [Wii]. Sega, Tokyo, Japan. Tokyo, Japan, Oct. 2009.

[423] Sega Sports R&D and Racjin. Mario and Sonic at the Winter Olympic Games. Game [DS]. Sega, Tokyo, Japan. Tokyo, Japan, Oct. 2009.

[436] Sleeping Beast Games. Spaceteam. Game [Android]. Sleeping Beast Games, Montreal, Canada. Montreal, Canada, Dec. 2012.

[440] Snap Inc. Snapchat Snappables. Game [Android]. Snap Inc, California, USA. California, USA, Apr. 2018.

[446] Square Enix. Final Fantasy XII. Game [PS2]. Square Enix, Tokyo, Japan. Tokyo, Japan, Oct. 2006.

[447] Square Product and Development Division 1. Final Fantasy X. Game [PS2]. Square Electronic Arts, California, USA. California, USA, Dec. 2001.

[466] Team Bondi. L.A. Noire. Game [Xbox 360]. Rockstar Games, New York, Games. New York, USA, May 2011.

[487] Ubisoft Montreal. Shaun White Skateboarding. Game [Wii]. Ubisoft, Rennes, France. Rennes, France, Oct. 2010.

[488] Ubisoft Paris, and Ubisoft Milan. Just Dance. Game [Wii]. Ubisoft, Rennes, France. Rennes, France, Nov. 2009.

[491] Valve. Half-Life 2. Game [Windows]. Valve, Washington, USA. Washington, USA, Nov. 2004.

[492] Valve. Portal. Game [Windows]. Valve, Washington, USA. Washington, USA, Oct. 2007.

[493] Valve South. Left 4 Dead. Game [Windows]. Valve, Washington, USA. Washington, USA, Nov. 2008.

[528] Witch Brema. Unpacking. Game [PC]. Humble Bundle, California, USA. California, USA, Nov. 2021.

Software & Manuals

[207] SPSS Inc. *IBM SPSS Statistics for Windows, Version XX*. Armonk, New York, USA: IBM Corp., 2024.

[281] MathWorks. *Stepwise Regression Documentation*. URL: https://www.mathworks.com/help/stats/stepwiselm.html.

[282] MATLAB. *Statistics and Machine Learning Toolbox v24.1*. Natick, Massachusetts, USA: The MathWorks Inc., 2024.

[283] MATLAB. *version 24.1.0.2537033 (R2024a*. Natick, Massachusetts, USA: The MathWorks Inc., 2024.

[309] Shane T Mueller. *The PEBL Manual*. 2nd ed. pebl.sourceforge.net, 2018.

[310] Shane T Mueller. *The Psychology Experiment Building Language (Version 2.0)*. [Software]. 2018. URL: https://pebl.sourceforge.net/.

[313] ST Mueller. "The PEBL pursuit rotor task". In: *Computer Software retrieved from http://pebl. sourceforge. net* (2012).

[504] Daniel Vasilaky. *vif(X)*. 2024. URL: https://www.mathworks.com/matlabcentral/fileexchange/60551-vif-x.

Part V
Appendices

The following chapters contain additional information to support the thesis. Each chapter is self-contained.

Appendix A

Ability Battery Minigames Designs

This chapter covers the design of the mini-games for the ability battery. These minigames are the tasks of the player profiling tool. They are informed by the ability tests from Ch. 10. We further explore inspirations from commercial minigames that seem to implement similar ability tests before explaining the game designs.

A.1 Minigame Inspirations

This section collects examples of mini-games that seem to match the structure of various cognitive and motor ability tests. They serve as inspiration and small case studies for ways we can design a game for our experimental purposes.

A.1.1 Finger Tapping Test

Mario Party 5's Will Flower minigame (Fig. A.4) [197] is an example of a game that mimics the structure of a FTT. It tasks players with repeatedly pressing the A button to revive a flower; players compete against each other to do this in a set time limit (the game times out at 5 minute). We could modify this gameplay to be a more suitable FTT by reducing the time limit to 10 seconds (to match existing FTT structure). The design choice to be made here is whether there is a set number of button presses that we are expecting the player to achieve (and so animation of the flower would be tied to a number), or are we looking to

Figure A.4: Will Flower, from Mario Party 5 [197] as an example of a game that could double as a Finger Tapping Test.

max out the number of presses the
player will try to do in a session
(and so have the animation move at a preset pace and just record how quickly they press). Since we are not looking to identify the maximum human limits or find a normative number, it makes sense for our performance focused design to have a normative number of presses that we design the challenge for, and then we can run the participant through several iterations based around that normative number.

A.1.2 Simple and Parametric GNGs, and Stroop Colour Word Test

A reminder that the difference between simple and parametric GNGs are their complexities. Simple GNG require straight forward responses (including potentially incongruent responses) to stimuli, where parametric GNGs add cognitive load to the task by asking the participant to interpret when the stimuli appears. With this in mind most simple mini-games (and thereby most examples) will be of simple GNGs.

Figure A.5: Video clip of Shy Guy Says mini-game from Mario Party[193]. Click the image to play.

Shy Guy Says (Fig. A.5) [193, 194, 323] mimics the structure of a Simple GNG. While the aesthetics are different (taking place on a pirate ship in Mario Party 1 and Superstars, and in the air for Mario Party 2) the core game is the same; the Shy Guy in the centre of the screen will raise a flag, players must then raise the same flag. Players who raise the incorrect flag or do not raise a flag are eliminated from the game. The last player standing wins; draws are possible if the last players are eliminated on the same round. What makes this a good GNG analog is how the responses are related to the stimuli, and there are many repetitive trials (in comparison to a reaction time task where it would be solely based on who responded first, not on who responded correctly). In order to modify this game for testing, the player cannot be eliminated with a wrong response — ergo we would need to create some other type of disincentive to keep the game feeling meaningful.

In Looking for Love (Fig. A.6) [324] players press directional input matching the location of the heart on the screen. The heart can appear in one of four locations (top, bottom, left, or right of the screen). This fits the basic parameters of a simple GNG (as stimulus appears perform response). This version adds two levels of difficulty; firstly, it provides distractor stimuli in the form of other icons appearing alongside the heart. Secondly, like a Stroop

test it changes the colour of the heart icon between trials. Initially the heart is coloured red (matching what participants expect a heart to be) and then over the trials it can be coloured other colours. Similarly, the distractor icons can be coloured either the same or different colours than the heart to try and further confuse/distract the players. The mini-game is not timed; players must play through 10 rounds (i.e. trials). Scoring is based on speed of correct response (first person to look gets 5 points, with subsequent participants getting a reduced score).

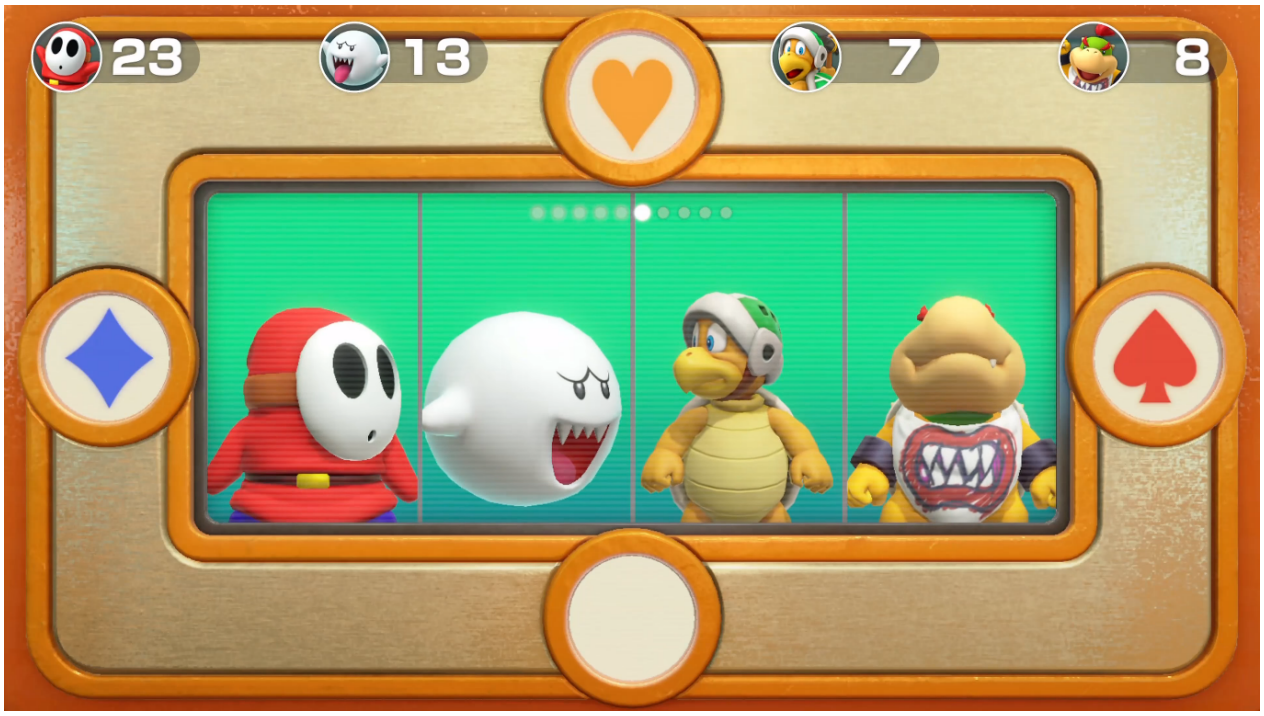


Figure A.6: Video clip of Looking for Love mini-game from Super Mario Party[324]. Click the image to play.

Similarly, Don't Look (Fig. A.7) [322] has players react with a directional input to a stimulus. Here the players are shown an arrow on screen pointing in a cardinal direction; the players must respond by looking in any direction except the one that matches the arrow. This is an in-congruent response condition for the GNG test. There are 10 rounds; for rounds 1 and 2 only 1 arrow is shown, 2 arrows are shown for rounds 3 to 6, and 3 are shown for 7 to 10. Each round has a time limit in which the players inputs are registered; this time limit is reduced for every trial.

Thundering Dynamo (Fig. A.8) [334] has the player's response be button mashing based on the colour shown on screen. When the stimuli is green the player must respond by mashing the B button, and when it is blue they must mash the A button. The game ends when one of the player has filled up their energy bar from button mashing; in this way there are no clear trials. As well, because the response is button mashing instead of a single press, the stimuli will always change to the other colour leaving the unknown variable to be when and for how long will it change.

Figure A.7: Video clip of Don't Look mini-game from Mario Party 9[322]. Click the image to play.

A.1.3 Object Categorization and Identification Tests

We start by looking at pure identification tasks.

Absent Minded (Fig. A.9) [324] works like an object identification task. The game is played across 3 rounds, each lasting at most 15 seconds. Players are shown 8 images, and are given 3 options at the bottom of the screen. The player must select which of the 3 options does not exist in the set of 8 images. The difficulty (and ability test) comes from how the 8 images are presented. In the first round the images are flashed one at a time on screen in a random order. In the second round, the images are heavily pixelated and become more focused over the course of the round. The third round cuts the image into tiles and places the tiles randomly across the screen meaning that the player cannot see a

Figure A.9: Second round of Absent Minded mini-game from Super Mario Party[324], where the images are pixelated. Click the image to play video clip of game.

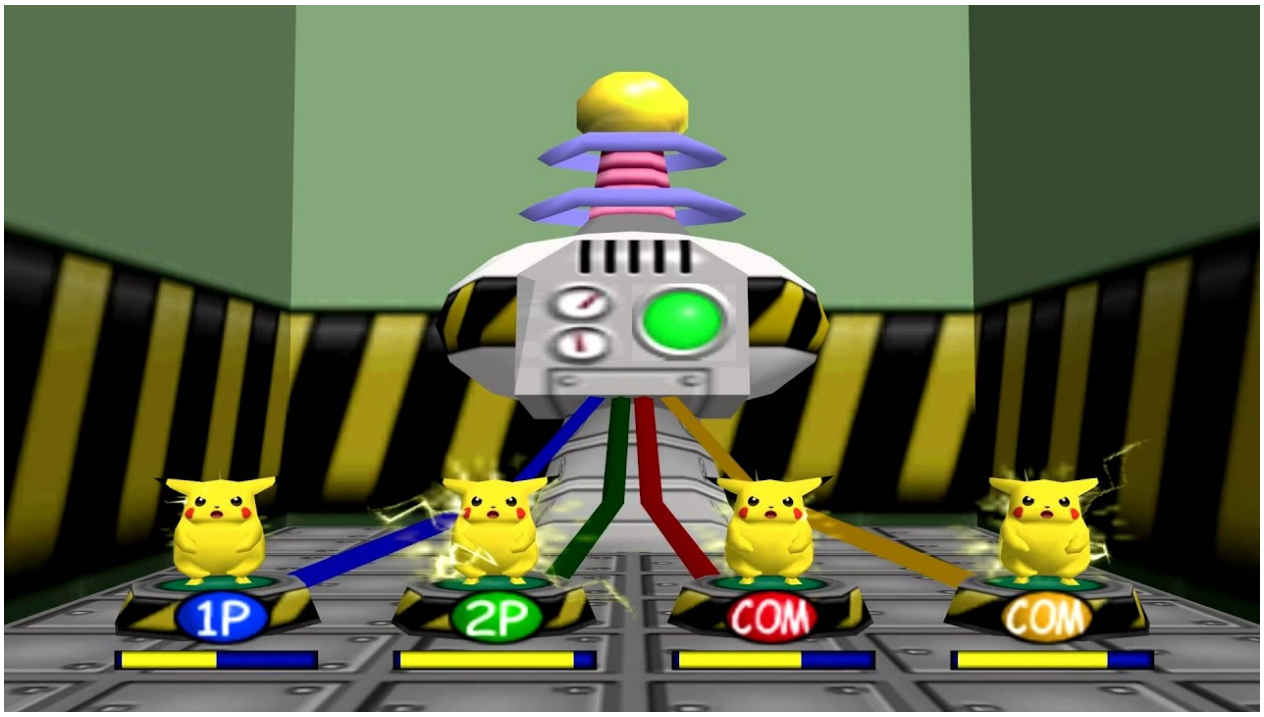


Figure A.8: Video clip of Thundering Dynamo mini-game from Pokemon Stadium[334]. Click the image to play.

complete image of any object. Players are scored based on how quickly they can correctly identify the missing image (fastest player receives 5 points, then 3, then 2, then 1). The individual rounds mimic conditions like the Shape Detection Task, Incomplete Letters Task, and Progressive Silhouettes Task. The structure of test is good, but would need to be adapted to single player set ups, and potentially include more rounds for trials.

Figure A.10: Crowd Cover from Mario Party 3[195]. Click the image to play video clip of game (video contains commentary unassociated with this thesis).

Crowd Cover (Fig. A.10) [195] presents the players with a picture which is obscured by moving characters. The players have 30 seconds to decide which of the 3 options the game presents matches the obscured figure. Over the duration of the mini-game the crowd decreases in size, making it easier to see the figure underneath. The difficulty here is determined by how different the options are and how many characters are covering the image.

We now look at object categorization tasks. These are more common in mini-games and combine object identification as the initial step, and categorization as the secondary step for these games.

Sort of Fun (Fig. A.11), from Super Mario Party [324], has players sorting types of balls into baskets, mimicking categorization and identification

tasks. Players have 30 seconds to work together to pass the balls to the person stationed at the correct basket; players can either pass the ball left or right, or drop the ball in the basket in front of them. For ever correctly sorted ball, the players get one point; every incorrectly sorted ball loses one point. Players are graded by their overall team score (with ≥ 10 being the top score). The game has a two step process (identifying if the ball matches your basket, identifying which direction the ball should move). Balls are a token change (they are all the same size, shape, and type with just their visual patterns being different) making it a within-category identification task. Every ball counts as a trial, making the whole experiment 30 seconds; the only thing that would need adjustment is making it a solo experience. The design decisions with that are how to adjust the inputs when you're not just choosing who to pass the ball to.

Figure A.11: Sort of Fun, from Super Mario Party [324] as an example of a game that could double as an Object Identification Test.

Figure A.12: Mail Sorting mini-game from The Legend of Zelda: The Wind Waker. Click the image to play video clip.

The mail sorting mini-game in The Legend of Zelda: The Wind Waker (Fig. A.12) [327] also acts as an object identification/-categorization task. Players are given 30 seconds to sort as many letters as possible based on the symbol on the letter. There are 6 different symbols, and the player gets 1 rupee (in game currency) for every letter sorted correctly; the reward increases based on the number of sorted letters (2 rupees per

letter past 20, 3 rupees per letter past 25). Every letter is a trial and the rewards increase based on performance. The choice to set goals at 20 and 25 letters shows an interesting assessment of what is possible yet difficult to accomplish.

A.1.4 Change Detection Tests

Odd Card Out (Fig. A.13) [198] tasks players to identify which of the three cards presented is different from the others. The game has 4 characters it could show for the test; each character has 5 unique tokens (Fig. A.14). In every round, a character is chosen by the game and then two cards from that character's set are selected (one as the target card, the other as a the comparison cards) to make the 3 presented to the player. Players have 5 seconds to identify which card is different. The first person to identify it gets the point. Once a player reaches two correct answers they win the mini-game. Odd Card Out is a comparison task, but does not exactly mimic a canonical CDT because all the tokens are shown at the same time. To use this as an effective assessment, we need to understand the magnitude of differences between the images.

Figure A.13: Video clip of Odd Card Out minigame from Mario Party 6 [198]. Click the image to play.

Figure A.14: Different token cards for Odd Card Out.

A.1.5 Serial Reaction Time Tasks (SRTT)

Recall SRTT are almost mechanically identical to GNG because they just ask players to respond correctly to different stimuli multiple times in a row. To this end, the examples here could easily be used as GNG examples as well.

A.2 Minigame Designs

All of the designed ability battery mini-games are created in the Unity engine. The mini-games take a JSON file for adjusting the variables and set-up of the ability battery. This allows for easy adjustment of the battery without needing to manipulate code, which makes it more portable for other researchers to use. The mini-games output the raw event data from the game to a JSON file. This data needs to be cleaned to be interpreted easily.

It is important to note that we will only be covering the mini-games that are useful measures for abilities we suspect are involved in Button Mashing challenges. We do this to keep the thesis concise and focused. However, more mini-games were developed that measure abilities outside of button mashing challenges as well.

A Quick Aside...

As a reminder, the design and coding of these games was done by various other students under my supervision. The code architecture, aesthetic choices, and details are their intellectual work. In their supervision I provided the guiding criteria for development (i.e. abilities they should be focusing on, general type of output needed) and acted as producer/project manager on their games (e.g. evaluating designs, assisting with decisions, helping with inspiration from other games). My intellectual contribution to this work is the idea, goals, and then the subsequent experiment design that these mini-games are used in.

A.2.1 Digger: Finger Tapping Proxy

Designed and coded by Mactivision student group as part of their capstone project.

Gameplay Description

The *Digger* mini-game presents tasks the player with digging from the surface of the Earth to a treasure buried underground within a set amount of time (default 100 seconds). To do this players mash a specified button (default 'B' button) as quickly as they can. Every button press moves the player down a specific amount — hence there is a specific number of presses that the player needs to reach the treasure (default 100).

Abilities Measured and Relationship to Existing Tests

The Digger game is designed to measure finger pressing, and so is based on and compared with the Finger Tapping Test (FTT). The FTT measures the number of presses that can

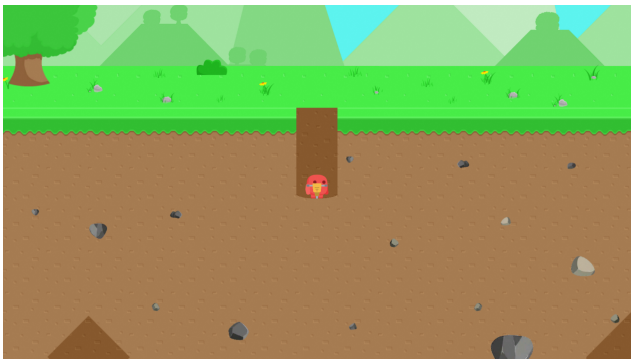


Figure A.15: Digger Mini-Game. Tests player finger pressing.

be made in 10 seconds; here we similarly set a time limit and measure number of presses. The difference here is that we set a "goal" number of presses that should be made in the 10 seconds. We found that able-bodied cis-men between the ages of 20-28, average approximately 60 presses per 10 seconds [32]. As we look for various normative data sets, we use this as a (albeit problematically skewed) benchmark for "average" performance.

Variable Components and Difficulty

Digger allows the tester to vary the following components:

Variable	Explanation	Effect on Difficulty	Variable Name	Default
Game length	The amount of time the game goes for before ending the instance.	Time restriction limits the total number of presses possible, and makes the speed of the pressing action the main thing being tested.	MaxGameTime	100
Goal presses	The amount of presses the player must make to successfully complete the game.	The goal number scales the difficulty in relation to the game length.	DigAmount	100
Key	The specific keyboard key whose presses are counted towards the goal.	The size and location of the key will make certain buttons easier or more difficult to press.	DigKey	B

For the purpose of measuring abilities, we will hold the game length and key to be fixed for the study. Based on the information we have on the FTT, we will hold game length at 10 seconds since that's the original metric for the analog FTT. We leave *keyCode* at its default setting. Therefore the only changing variable is the goal number.

For testing with this mini-game we believe that a "medium" level of difficulty would be a target goal of 60 presses per 10 seconds, which is in line with the found average [32]. A

lower number would create an "easy" variant, and a higher number would create a "harder" variant as fatigue sets in. The same study reports the changes in intertap interval over the number of button presses and uses it to evaluate when fatigue starts to have an observable effect on performance. They specifically note that fatigue begins to affect the performance of their participants around the midpoint of the task (4-5 seconds, or approximately 30-35 presses), and that post 60 presses the intertap intervals were no longer tightly coupled, with the graph maxing out at 81 presses [32]. Using these as our guidelines we set the "easy" goal to 35 presses, and the "hard" goal to 80 presses.

Measurements/Outputs

The Digger file is looking to measure the button presses that happen during the mini-game. It catalogues every button press event and tells us which button is pressed, and when. The output is a series of JSON entries which look like Listing 1.

```
{
    "keyCode": 98,
    "keyDown" : true,
    "eventTime" : "yyyy-mm-ddThh:mm:ss"
}
```

Listing 1: Example output of Digger mini-game.

The total number of events logged lets us know how many total presses there are. In this way we can see whether a participant has met the goal (number of presses), if they have made any errors (which button is pressed via *keyCode*), and how quickly they press (based on difference between *eventTime*).

A.2.2 Stage: Token Change Detection Proxy

Designed and coded by Vansh Pahuja as part of their M.Eng Project.

Gameplay Description

The *Stage* mini-game tasks the player with identifying if there have been any changes made to the on-stage performers. The player is shown the first set of performers walking onto the stage, and after a short pause on stage they walk off. The second set of performers then walks on the stage, and the player must decide whether anything has changed by pressing one button for yes (default Right shift) and another button for no (default Left shift).

Abilities Measured and Relationship to Existing Tests

The Stage game is designed to measure token change detection, and is based on the canonical change detection task (CanonCDT). The CanonCDT measures whether a person can accurately determine whether there has been a change in the number or colour of squares presented to them. The basis of this test is that the array of items is stored in the player's

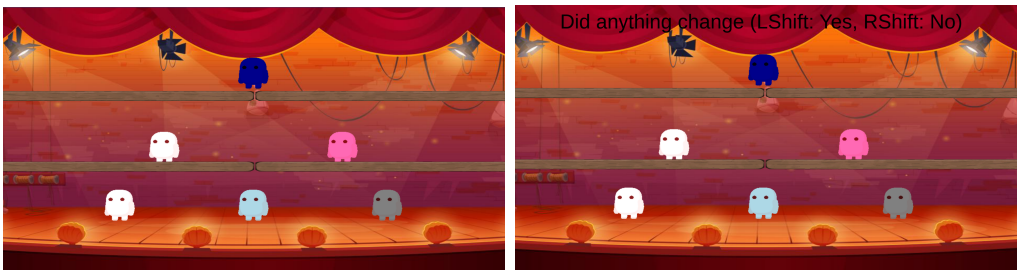


Figure A.16: Stage Mini-Game. Tests player's token change detection. Left image is the original set, right image is the second set with the instructional prompt.

working memory, and so long as the number of stimuli presented is less than or equal to the working memory capacity the player should be able to easily identify if there are any changes. If the number of stimuli exceeds the working memory capacity, the probability of an incorrect response (i.e. saying no change when there is a change, or change when there is no change) increases. While Miller's original "magic number" is 7 ± 2 [299], the psychology community generally believes that the working memory capacity of perfectly encoded information is between 2 and 5 objects, depending on their complexity [35][1]. Therefore difficulty should be scaled around this understanding.

Variable Components and Difficulty

Stage allows the tester to vary the following components:

For simplicity purposes we only modify the difficulty level during testing. The difficulty variable changes the total number of stimuli on the screen. Difficulty level 1 displays 3 stimuli, which is solidly between the 2 to 5 objects that we generally accept as allowing for perfect encoding. Difficulty level 2 displays 6, which falls slightly outside the 2 to 5 objects, but still remains in the "magic number". Difficulty level 3 displays 9 stimuli, which is the high end of the "magic number" and squarely outside of the 2 to 5 object range. The game length will be set to 500 seconds for each instance; this is to account for animation time, and waiting time for the player to input their answer. This time limit allows for each instance to have approximately 5 trials per instance, accounting for animation time and wait times for the player to make their decision. While the original CanonCDT did not impose time limits on the participants, we believe that in order to keep the ludic nature of these games in the battery these constraints should remain. We have set the game to not timeout if the player is currently making a choice, so that players do not interpret the mini-game to be unfair for cutting them off before they made a choice.

[1] Just a reminder, that we are subscribing to the discrete model of working memory capacity as that is the paradigm the change detection tasks have been developed under. There is mounting evidence based on how cueing influences working memory that this discrete model may not be sufficient [35]. There is also a competing resource paradigm for working memory that instead proposes that every stimuli seen is attended, but the quality of information encoded in memory about them is reduced as the number of stimuli increases [137].

Variable	Explanation	Effect on Difficulty	Variable Name	Default
Game length	The amount of time the game goes for before ending the instance.	Time restriction limits the total number of presses possible, and makes the speed of the pressing action the main thing being tested.	MaxGameTime	90
Number of trials	The total number of trials in the instance.	–	MaxPrompts	5
Number of stimuli	The maximum number of stimuli that can be on screen.	The more stimuli the hard it is to identify changes because you have to remember more.	MaxPlayersDisplayed	10
Difficulty level	Controls the number of stimuli on screen for different levels.	More stimuli means harder difficulty because there is more to remember. Diff 1 is 3 stimuli; Diff 2 is 6 stimuli; Diff 3 is 9 stimuli	Diff	2

Measurements/Outputs

The performance measurement for Stage is the number of correctly identified changed scenes. Stage catalogues every user input during the "prompt" section (as seen in right image of Fig. A.16). The output is a series of JSON entries which look like example 2.

The response time for each trial is the difference between the *choiceTime* (when the input is made) and *eventTime* (when the second set is shown to the player) variables. The *colorChanged* variable indicates whether there was a change (true) or not (false). When compared with the *choice* variable it lets us know whether the player successfully identified the change (RShift) or not (LShift). The output also provides additional information about the colour order of the stimuli (*colors1...9*[2]) and what the original stimuli color (*colorOriginal*) was and the changed colour (*colorNew*). These pieces of additional information can be used to understand the type of error (e.g. if the colours were similar to each other).

A.2.3 Looking: Selective Attention and Inhibition Proxy

Designed and coded by Vansh Pahuja as part of their M.Eng Project.

Gameplay Description

The *Looking* mini-game tasks the players with finding the object that matches the given stimulus. Players are presented with a target item (Fig. A.17, left) and then are shown four items on screen (Fig. A.17, right). If the item is on the screen, the player must press its

[2]Details about the enumerated colour values will be found in Vansh's thesis.

```
{
    "color1": 3,
    "color2": 4,
    "color3": 1,
    "color4": -1,
    "color5": -1,
    "color6": -1,
    "color7": -1,
    "color8": -1,
    "color9": -1,
    "colorsShown": null,
    "correct": true,
    "choice": "RightShift",
    "colorChanged": true,
    "colorOriginal": 1,
    "colorNew": 6,
    "choiceTime": "2023-07-05T11:04:19.5455543-04:00",
    "eventTime": "2023-07-05T11:04:16.5650428-04:00"
}
```


Listing 2: Example output of Stage mini-game.

corresponding arrow key to identify it. If the item is not on screen, the player must press 'X' to indicate that it is not there. The game is timed and players are incentivized to go quickly.

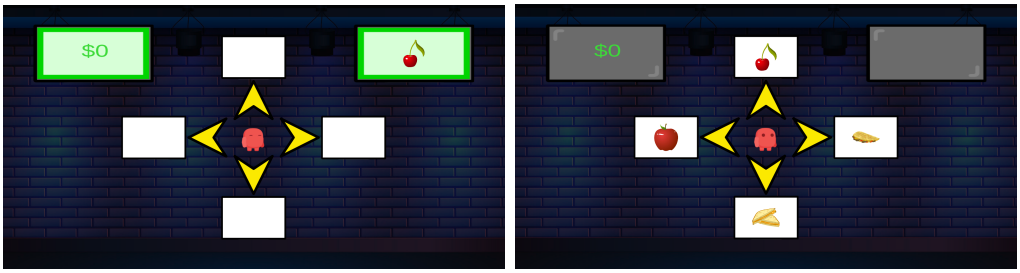


Figure A.17: Looking Mini-Game. Tests selective attention and inhibition. Left image shows target stimulus. Right image shows presented options.

Abilities Measured and Relationship to Existing Tests

The Looking mini-game is designed to measure selective attention and inhibition via choice reaction task (Go/No-go paradigm). Players are required to prioritize speed in this test similar to the Looking for Love mini-game in Super Mario Party. The item set has many overlapping features that could make quick object identification difficult; for example there are two drinks in a martini glass with their main difference being colour (red and orange)[3].

[3]The quantification of item similarity is displayed in a visual similarity matrix in Vansh's thesis.

While this task also uses object recognition abilities, the focus on speed and the overlapping features between items allows us to target inhibition and selective attention because it requires significant restraint to not just go with the first seemingly correct item, especially in cases where the target is not actually present. Normative performance data does not exist (and would be somewhat meaningless) for Go/No-go paradigm tasks; as such while we care about the difficulty of the tasks, we do not base the difficulty adjustments around any normative baselines but rather our understanding of how the variable elements affect task performance more holistically.

Variable Components and Difficulty

Looking allows the tester to vary the following components:

Variable	Explanation	Effect on Difficulty	Variable Name	Default
Game length	The amount of time the game goes for before ending the instance.	Time restriction limits the total number of presses possible, and makes the speed of the pressing action the main thing being tested.	MaxGameTime	90
Number of trials	The total number of trials in the instance.	–	MaxFoodDisplayed	15
Number of stimuli	The total number of unique stimuli.	More stimuli means there are more to distinguish between for sorting; fewer stimuli types means that players can figure out categorization through trial and error.	UniqueObjects	6
Number of stimuli before new target	Average number of stimuli dispensed before it presents you with a new set.	Player must pay attention and update their working memory to make sure they are matching against the currently active set.	AverageUpdateFrequency	3
Variance in updating	Variance of avgUpdateFreq	—	UpdateFreqVariance	0.3

For the purposes of our work we set the number of trials for an instance (MaxFoodDisplayed) at 18. Since speed is a factor of our work, we maintain the default time limit for this instance. If the player is unable to complete the total number of trials for an instance in that time, the missed trials will count as failures.

While the intrinsic difficulty of each choie is set based on the qualities of the items shown, we can control some difficulty through the update frequency of the target stimulus. The more frequently the target changes, the more likely the player is to make errors because

they're remembering the previous target. Simple GNG do not do this, however Parametric GNG will often provide multiple rules that change response behaviour — we believe this target updating feature is in line with Parametric GNG design. We change the number of stimuli between target changes (AverageUpdateFrequency) to reflect the instance's difficulty. Easy instances of Looking will have an AverageUpdateFrequency equal to the number of trials (ergo only one target stimulus like a Simple GNG). Medium instances will have an AverageUpdateFrequency of 6, ergo creating three stimulus blocks. Difficult instances will have an AverageUpdateFrequency of 3, creating 6 stimulus blocks.

Measurements/Outputs

The performance measurement for Looking is based on accuracy (correct responses to incorrect responses) and response time measured in milliseconds. The output is a series of JSON entries which look like example 3. For every trial *correct* tells us whether the response was


```
{
  "_goodObject": "cosmopolitan",
    "correct": "True",
    "objectsShown": [
        "broccoli",
        "cosmopolitan",
        "frittata",
        "apple"
    ],
    "choice": "Right",
    "choiceTime": "2023-06-08T11:38:14.3970855-04:00",
    "eventTime": "2023-06-08T11:38:13.8642335-04:00"
}
```


Listing 3: Example output of Looking mini-game.

correct (true) or not (false). We can use the total count of trials, and the number of correct and incorrect responses to calculate the accuracy of the player. As well, we have the difference between *eventTime* and *choiceTime* which gives us the response time for each trial. We also output the additional information about which objects are shown (*objectsShown*), what the target object is (*_goodObject*) and which button was pressed. We can use this additional data to consider reasons why incorrect responses may have happened.

A.2.4 Cake: Object Recognition (Categorization) Proxy

Designed and coded by Vansh Pahuja as part of their M.Eng Project.

Gameplay Description

The *Cake* mini-game tasks the players with sorting various food objects into three categories: desserts, meals, or fruits/vegetables. The stimulus flows along a conveyor belt and the player

must move the sorting bins up and/or down until the correct box lines up with the conveyor belt.

Figure A.18: Cake Mini-Game.

Abilities Measured and Relationship to Existing Tests

The Cake mini-game is designed to measure object recognition; it is based on object categorization tasks and mini-games like Sort of Fun in Super Mario Party and the letter mini-game in Legend of Zelda The Windwaker. In Cake, food travels down a conveyor belt at a fixed speed, and players must sort the food into one of three categories (dessert, meals, fruits/vegetables). Inherently this relies on North American cultural understandings of the categories. As with the previous mini-game, normative data does not exist, nor would it make sense, for this type of task. As such when we discuss the difficulty of the mini-game and the ways we will adjust it, those decisions are based on our understanding of how the ability works and how the variables will affect performance in the task.

Variable Components and Difficulty

Cake allows the tester to vary the following components:

Difficulty in object categorization comes in two forms: the clarity of the object being categorized (i.e. it fits into a single category), and the time limit for making your decision. For the North American context of our experiment, the objects used were selected so they only fit into one of the three categories. The time limit for making decisions relies on the number of stimuli on the conveyor belt (AverageDispenseFrequency) and how quickly they reach the sorting bins (FoodVelocity). The higher the FoodVelocity, the less time the player has to decide on the object category and adjust the sorting bins. This makes each sorting task more difficult. This difficulty can be compounded by having the number of objects on the conveyor belt at one time increased (which is done by decreasing the AverageDispenseFrequency). If only one object is on the conveyor belt at a time, the player has time proportional to FoodVelocity to react. If there are multiple the perceived time to react is lower because subsequent objects will only be responded to after the first is dealt with (hence losing time in processing, categorizing, and reacting).

To this end, our "easy" version of this mini-game will have FoodVelocity 2.5 and reduce the AverageDispenseFrequency to 2. We increase the difficulty in the medium version by increasing the FoodVelocity to 3, and decreasing AverageDispenseFrequency 1.5. For the

Variable	Explanation	Effect on Difficulty	Variable Name	Default
Game length	The amount of time the game goes for before ending the instance.	Time restriction limits the total number of presses possible, and makes the speed of the pressing action the main thing being tested.	MaxGameTime	90
Number of trials	The total number of trials in the instance.	–	MaxFoodDispensed	20
Number of stimuli	The total number of unique stimuli.	More stimuli means there are more to distinguish between for sorting; fewer stimuli types means that players can figure out categorization through trial and error.	UniqueFoods	9
Number of stimuli at once	Average number of stimuli on conveyor belt between the food updates.	Every stimuli on the belt acts as a distractor for the current one to be sorted.	AverageDispenseFrequency	3
Conveyor belt speed	How quickly the stimuli move across the conveyor belt; variance of avgUpdateFreq	The faster the stimuli need to be sorted, the more pressure the player will feel and the less time they have to think.	FoodVelocity	2.25

"hard" version we increase the FoodVelocity to 3.5, and increase the the AverageDispense-Frequency to 1. We arrived at these numbers through play-testing the game while tweaking these numbers in order to experience when the mini-game felt noticeably more challenging.

Measurements/Outputs

The performance measurement for Cake is the accuracy of the player. The output is a series of JSON entries which look like example 4.

For every trial *correct* lets us know whether the response was correct (true) or incorrect (false). The additional information such as *objectType*, *_object* and *boxChoice* provide us with insight into incorrect trials. For example, we can compare *_object* to the *boxChoice* and see whether there is any cultural constraints that may have created the error (e.g. fruit like oranges being thought of as dessert may lead to incorrect labelling due to cultural differences).

As well we collect the *choiceTime* and *eventTime*, whose difference lets us review how much time the stimulus was on screen — and by extension how long the player could have had to process it.

```
{
    "objectType": 1,
    "_object": "cookie_chocolate_chip(Clone)",
    "boxChoice": 1,
    "correct": true,
    "choiceTime": "2023-06-08T11:56:36.6988288-04:00",
    "eventTime": "2023-06-08T11:56:33.212261-04:00"
}
```


Listing 4: Example output of Cake mini-game.

A.2.5 Recipe: Object Recognition (Identification) Proxy

Designed and coded by Vansh Pahuja as part of their M.Eng Project.

Gameplay Description

The *Recipe* mini-game tasks the players with recognizing whether a pair of sweets are the same or different from a target pair. The player is first presented with a target set of sweets (left Fig. A.19). The dispenser then begins to dispense two sweets at a time (Fig. A.19 right). If the pair matches the target set, then the player should pack it for shipping (right arrow). If it does not match, they should dispose of it (left arrow). Once they've made their decision, a new pair of sweets will be dispensed to be identified as either the same or different from the target set. The target set will change based on an update frequency, meaning that the player will also need to be aware of what the current target set is.

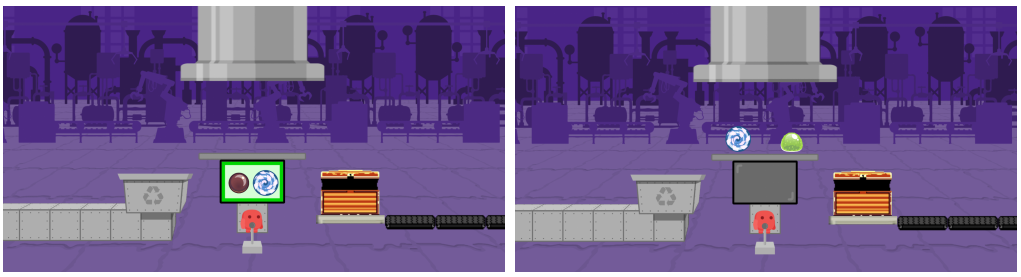


Figure A.19: Recipe Mini-Game. Tests player's object recognition (identification). Left image is the target pair, right image is a trial the player must sort.

Abilities Measured and Relationship to Existing Tests

The Recipe mini-game is designed to measure object recognition, and is based on object identification tasks. This mini-game shares similarities to *Looking*. Players are prioritizing speed, and dealing with an object set where there are overlapping features meaning there is a non-trivial amount of attention that is needed to complete this task. These similarities show that it is a simpler choice reaction task (or a simple Go/No-go) and so is also measuring

inhibition and selective attention at the same time. We have emphasized the object identification element of this mini-game through the use of object pairs; since the player needs to see whether both stimuli match the "recipe" (target stimuli) the identification aspect of this game should be more involved than the regular choice reaction tasks which focus on selective attention/inhibition.

As with object categorization and selective attention/inhibition, normative data does not exist, nor would it make sense, for this type of test. As such when we discuss the difficulty and ways we modify it those decisions will be based on our understanding of how the ability works and how we can affect the performance of the game.

Variable Components and Difficulty

Recipe allows the tester to vary the following components:

Variable	Explanation	Effect on Difficulty	Variable Name	Default
Game length	The amount of time the game goes for before ending the instance.	Time restriction limits the total number of presses possible, and makes the speed of the pressing action the main thing being tested.	MaxGameTime	90
Number of trials	The total number of trials in the instance.	–	MaxFoodDispensed	20
Number of stimuli	The total number of unique stimuli.	More stimuli means there are more to distinguish between for sorting; fewer stimuli types means that players can figure out categorization through trial and error.	UniqueFoods	9
Number of stimuli before new target	Average number of stimuli dispensed before it presents you with a new set.	Player must pay attention and update their working memory to make sure they are matching against the currently active set.	AverageUpdateFrequency	3
Variance in updating	Variance of avgUpdateFreq	—	UpdateFreqVariance	0.3

As with object categorization difficulty is determined partly by the visual similarity between the objects[4], but also by the frequency of the target set changing. In this case, since our game is closer in design to *Looking*, we use the same variable values as Looking to determine our "easy", "medium", and "hard" variants. Each instance will be set to 18 trials

[4]The visual similarity between sweets is recorded in a visual similarity matrix in Vansh's thesis.

(MaxFoodDispensed), For the easy variant AverageUpdateFrequency will be equal to the number of trials (18) such that there are no changes. For the "medium" variant AverageUpdateFrequency will be 6, creating 3 blocks of trials. Finally, the "hard" variant will set AverageUpdateFrequency to 3, creating 6 blocks of trials.

Measurements/Outputs

The performance measurements for Recipe are the number of correct responses and the response time in milliseconds. The output is a series of JSON entries which look like example 5. For every trial we see what the "recipe" (target stimuli set) is, as well as the objects that


```
{
  "objectsSet": [
    "pinkstar",
    "greenjelly"
   ],
   "_object": [
    "greenjelly",
    "blueswirl"
   ],
   "choice": true,
   "correct": false,
   "choiceTime": "2023-06-08T11:41:37.0527443-04:00",
   "eventTime": "2023-06-08T11:41:35.9992559-04:00"
}
```


Listing 5: Example output of Recipe mini-game.

were dispensed. *choice* lets us know what the player's choice was for this trial (true or false about matching the recipe). We can compare this against *correct* which tells us whether the right answer was true or false. Counting the total number of correct choices for the test will allow us to calculate the accuracy of the player's identifications. Knowing the objects that were dispensed for each trial and what the target stimuli were will allow us to hypothesize about why errors may have occurred (e.g. in Listing 5 one of the target stimuli was there but not both). This can help us have more robust understandings about the performances of our players. We can also use *choiceTime* (when the player makes their decision) and *eventTime* (when the stimuli are dispensed) to find the response time for that trial. We can find the player's average response time to consider the speed-accuracy tradeoffs that may affect their overall performance score.

Appendix B

Correlational Study Extra Details

This chapter covers any details about the study design and procedures to ensure that another person could re-run this work.

B.1 Recruiting Participants

Participants are recruited by four processes:

- flyers posted on McMaster University's main campus,
- targeted email recruitment,
- social media posts, and
- snowball sampling.

A recruitment poster (Fig. B.20a) was made and submitted in accordance with the McMaster Research Ethics Board (MREB) guidelines. The recruitment poster was posted across campus in departmental buildings, libraries, and the student centre in order to ensure a wide net was cast as to potential participants.

A secondary poster (B.20b) was made to be attached to the email and social media posts. In this way consistent branding and information was provided to potential participants. E-mails were sent to undergraduate and graduate students in the Department of Computing and Software at McMaster University. A snowball recruitment script was prepared for asking participants whether they knew anyone else who would be eligible and interested in participating.

Participants are not being offered compensation for their participation in this study.

B.2 Pre-Study Survey

We collect information about the participant's demographics and gaming habits through a pre-session survey. Demographic data such as age, sex (assigned at birth), education level, and self-identified disability status are useful in explaining and comparing ability performance data against other normed data which often focuses on specific demographics. As well, this information will be useful when considering any significant deviations of performance between participants considering we do not impose significant restrictions on the potential participants. Information about game playing habits similarly contextualizes participant performance to some degree. We anticipate that those who play a variety of games much more frequently will perform better than someone who plays a single game only once a week due to greater game literacy (i.e. they understand the conventions of video games). Therefore any significant variations in performance could be explained by this difference in background knowledge between participants.

The following document is a copy of the survey participants must fill out.

Human Mechanics of Video Game Design

This study is being conducted by Sasha Soraine and Dr. Jacques Carette of the Department of Computing and Software at McMaster University. They can be reached via e-mail at sorainsm@mcmaster.ca and carette@mcmaster.ca respectively.

The purpose of the study is to confirm whether mini-games (short video games with clear goals) can be used to measure a player's performance for various human cognitive and motor abilities. Information gathered during this study will be written up as a doctoral thesis and used in conference and/or journal papers. People participating in this study must be between the ages of 18 and 64, and play video games (any genre, any platform) at least once a week.

This survey should take approximately 5-10 minutes to complete. Eligible participants will then be contacted to schedule two study sessions. The study sessions will take a total of 1 hour and 10 minutes to complete (50 minutes for the first session, 20 minutes for the second session), with a one-week gap between sessions.

To learn more about this study, particularly in terms of any risks or harms associated with the study, how confidentiality and anonymity will be handled, withdrawal procedures, incentives that are promised, how to obtain information about the study's results, how to find helpful resources should any questions or tasks make you uncomfortable or upset etc., please read the Letter of Information.

This study has been reviewed and cleared by the McMaster Research Ethics Board (MREB# 2347).

If you have any concerns or questions about your rights as a participant or about the way the study is being conducted, please contact:

McMaster Research Ethics Board Secretariat
Telephone: 1-(905) 525-9140 ext. 23142
E-mail: mreb@mcmaster.ca

* Required

1. Having read the previous preamble OR the linked Letter of Information, I understand that by clicking the "Yes" option below, I agree to take part in this study. *

   Mark only one oval.

   - ◯ YES, I agree to participate in this study. *Skip to question 2*
   - ◯ NO, I do not agree to participate in this study *Skip to section 7 (Withdrawing from Study)*

Consent Statements

This section outlines specific consent statements that must be answered before participating in the survey. Agreeing or disagreeing to these specific consent statements will not be used to disqualify participants from the study.

2. I agree to allow my study data to be stored and used for future research as described in the Letter of Information *

   Mark only one oval.

   - ◯ Yes
   - ◯ No

3. If I choose to quit the study, I agree to have my responses up to the point of quitting the study retained for use in the research. *

   Mark only one oval.

   - ◯ Yes
   - ◯ No

4. I agree to allow my anonymized study data be uploaded to an open science data sharing platform *

Mark only one oval.

- ◯ Yes
- ◯ No

Contact Information

The following information will be used to contact you in order to set up your study sessions and optionally to send you copies of the study results. This information will not be stored as part of the study data.

For contact purposes we require an e-mail from participants. Participants who prefer to coordinate over phone (call or text) may provide their contact number, but e-mail is the only required contact method.

Please note that your name and contact information will remain completely confidential and **will not be linked** with any of your study responses.

5. Please enter your full name and preferred pronouns. *

_____

6. Please enter your e-mail address. *

_____

7. What is your preferred methods of contact? Select all that apply.

Check all that apply.

- ☐ E-mail
- ☐ Phone (call)
- ☐ Phone (text)

8. Please enter your phone number (for calling).

_____

9. If it is different from the calling number, please enter your phone number for texting.

_____

Demographic Questions

These are questions regarding your personal demographics. They are necessary to help us understand differences in video game performance.

10. What is you age in years? *

_____

11. What was your sex assigned at birth? *

Mark only one oval.

- ◯ Prefer not to say
- ◯ Intersex
- ◯ Female
- ◯ Male

12. What is your highest level of completed education? *

Mark only one oval.

- ◯ No formal education
- ◯ Some elementary school (Kindergarten to Grade 8)
- ◯ Elementary School
- ◯ Some high school or equivalent
- ◯ High School Diploma OR GED
- ◯ Some college or university
- ◯ College or University
- ◯ Some post-graduate (Masters or PhD)
- ◯ Post-graduate (Masters or PhD)

13. Do you identify as a person with a disability that was present at birth, caused by an accident, or developed over time; *

• that encompasses any degree of physical disability, mental or developmental disability, sensory disability, learning disability, mental health / psychiatric disability, addiction, and life-threatening allergies;

• that may affect full participation in society (school / work);

• that may have been accommodated in workplace / school because of functional limitation as a result of the disability; or,

• who, as a result of self-perception, perception of others, environmental barriers, inaccessible attitudes, or a any combination of these factors, may experience unequal opportunity to access services by reason of the disability?

Mark only one oval.

- ◯ No
- ◯ Yes
- ◯ Prefer to self-identify
- ◯ Prefer not to answer

14. If you identify as having a disability, does your disability affect any of the following? Please select all that apply.

Check all that apply.

- [ ] Vision
- [ ] Hearing
- [ ] Fine motor control (e.g. fingers, toes)
- [ ] Gross motor control (e.g. legs, arms)
- [ ] Attention
- [ ] Short term memory
- [ ] Long term memory
- [ ] Executive functioning
- [ ] Other: _____

Video Game Playing Habits

These are questions about your video game playing habits and preferences. They are used to give us insight into how you approach the study.

15. Approximately how long have you been playing video games? *

Mark only one oval.

- ◯ Less than a year
- ◯ Between 1 and 4 years
- ◯ Between 5 to 9 years
- ◯ 10 or more years

16. On average, how many times a week do you play video games on any platform (e.g. phone, tablet, console, computer) *

Mark only one oval.

- ◯ Once a week
- ◯ 2 to 4 times a week
- ◯ 5 or more times a week

17. On average, how much time do you spend on a single play session (from turning the game on to turning it off)? *

Mark only one oval.

- ◯ 5 to 10 minutes
- ◯ 15 to 30 minutes
- ◯ 30 minutes to 1 hour
- ◯ 1 - 3 hours
- ◯ More than 3 hours

18. What are the games you play most frequently? *

_____

_____

_____

_____

_____

19. What platforms do you commonly use to play games? Please select all that apply. *

Check all that apply.

- [ ] Computer
- [ ] Smartphone or tablet
- [ ] Handheld console (e.g. Playstation Vita, Nintendo 3DS)
- [ ] Console (e.g. Xbox One, Playstation 5, Nintendo Switch)
- [ ] Other: _____

20. On a scale of 1 to 5, how comfortable are you with using a keyboard and mouse? *

Mark only one oval.

	Very uncomfortable
1	◯
2	◯
3	◯
4	◯
5	◯
	Very comfortable

21. From the list below, please select any video game controllers that you have used in the past. *

Check all that apply.

- [ ] Keyboard
- [ ] Mouse
- [ ] Joystick and buttons
- [ ] Gamepad (e.g. Xbox controller, Playstation controller)
- [ ] Motion controller (e.g. Wii Remote, Switch Joy-Cons, Playstation Move)
- [ ] Full motion controller (e.g. Kinect)

22. Do you consider yourself a gamer? *

Mark only one oval.

- ◯ Yes
- ◯ No
- ◯ Other: _____

Survey Complete!

Thank you for completing the pre-study questions. Your responses have been submitted. Our researchers will check your age and gameplaying habits to see if you are eligible for the study before contacting you to set up a session.

If you have any follow up questions please feel free to contact Sasha Soraine by phone (289-434-4053)or via e-mail (sorainsm@mcmaster.ca).

23. Please select any that apply to you:

Check all that apply.

- [ ] I would like a copy of the study results sent to the contact e-mail I provided.
- [ ] I would like to be contacted to participate in further studies for this research program.

Withdrawing from Study

Thank you for your time. You have decided to quit this study. None of your responses have been collected or stored.

(a) (b)

Figure B.20: MREB approved recruitment visual materials; B.20a is approved for physical posting, B.20b is approved for posting on social media.

B.3 Detailed Data Analysis - Accuracy

Since the Battery is intended for capturing abilities in a gaming context, we originally considered comparing the data by its accuracy. We believed focusing on accuracy over reaction time would be appropriate as games frequently use accuracy as a binary success condition, with reaction time being a secondary feature indicating level of success. However, raw accuracy measures are only reliable for showing individual differences when participants make sufficient errors, and the speed-accuracy trade-off is controlled for by fixing reaction times or making them irrelevant (e.g. 118). Since we recruited participants with gaming backgrounds, and modeled the Battery after existing games, it is possible that the spread of accuracy measures will be too close to be useful. As well, all tasks (except Battery's Cake game) incentivize participants to as accurate and as fast as possible. Since we cannot account for how that was interpreted, we may have participants who have lower accuracy scores due to prioritizing speed and higher scores from participants who prioritized accuracy. Comparing the data by accuracy measures showed data distributions and correlations in-line with these concerns.

This section takes a close look at the accuracy data for every pairing of minigame and PEBL task, to analyse validity and reliability.

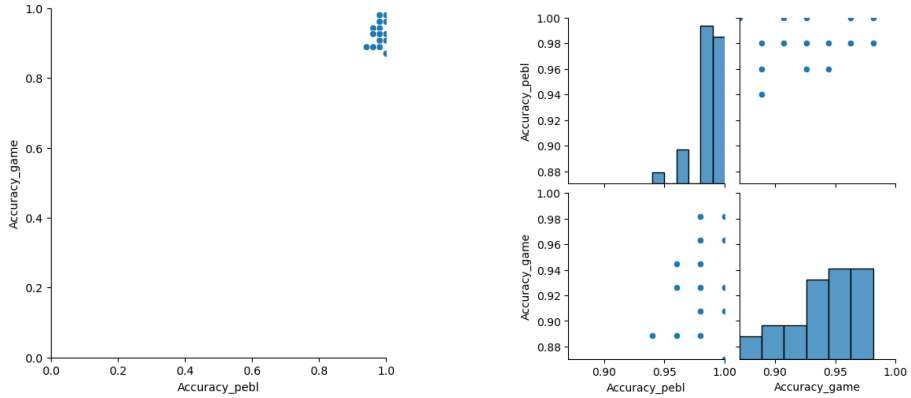
B.3.1 Validity

To better understand what is happening with this data, we look at each reported correlation through it's scatterplots. For validity testing, we always report PEBL (standard measure) in the X-axis, and the Battery results (new measure) in the Y-axis.

Digger-Tapping: The Digger-Tapping pair is measured by press rate (presses/second). Since this is the measure we stick with for study results, the details about this analysis are in B.4.1.

Looking-Choice: There was no reason to remove outliers in this data.

Fig. B.21 compares the accuracy scores for Looking and Choice Reaction Time. We see no significant correlation ($\rho = 0.214262, p = 0.242698571$; $r = 0.244763, p = 0.184485$). Zooming in on the data for Choice (Fig. B.21b) we see clear bucketing. PEBL scores only exist as either 0.94, 0.96, 0.98, or 1.00. The Battery has a more varied distribution, although still tightly in the 0.8 to 1.00 range.



(a) PEBL Choice Reaction Scores (Mean: 0.9852, s.d. = 0.0155) vs. Battery Looking Scores (Mean: 0.9397, s.d. = 0.0361), Accuracy.

(b) Zoomed in pair plots and distributions.

Figure B.21: Validity test for Looking mini-game and Four Choice Reaction Time PEBL Test.

At first we thought this may be a case of different intents between PEBL's Four Choice and Looking. Where Looking encourages speed through the score system, Four choice does not limit the amount of time participants have to make a decision about their input. If participants chose accuracy in their accuracy-speed trade-off, it may explain some of the bucketing but we would expect to see significantly longer reaction times. However when we look at the relationship between reaction time and accuracy for both Four Choice and Looking, they seem uncorrelated and PEBL's reaction times are shorter than Battery's (Fig. B.22).

Having thought more on the design of Four Choice Reaction and Looking, this may be a case where PEBL is "easier" than the Battery. The four choice task asks participants to

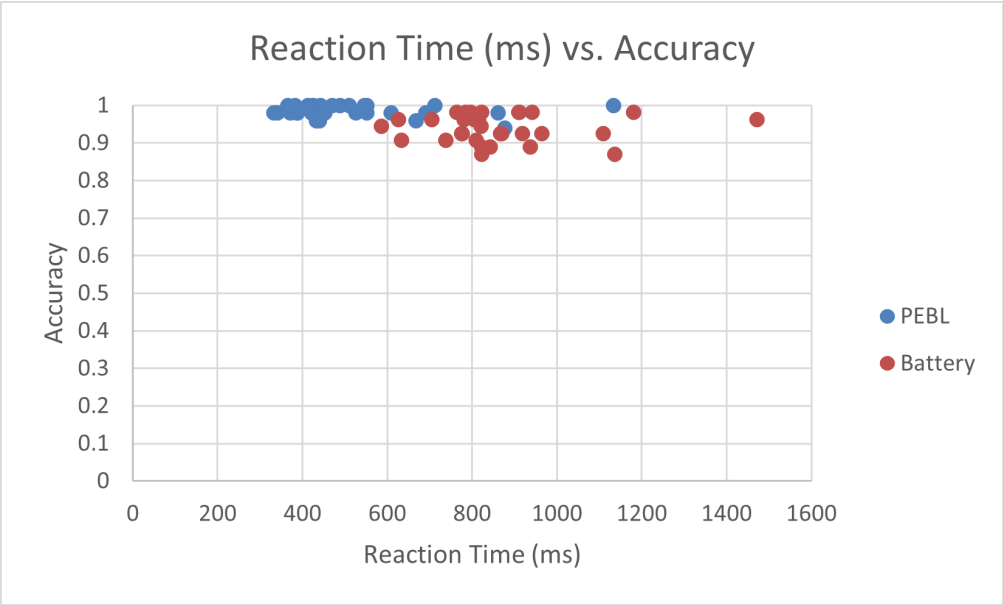


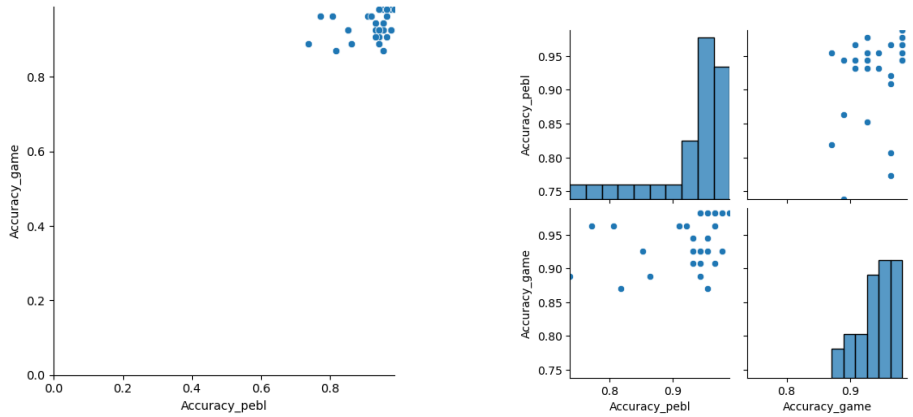
Figure B.22: Comparing Reaction Time (ms) to Accuracy Scores for both PEBL Four Choice Task and Battery's Looking Game. There is no statistical correlation between response time and accuracy in this data.

select the screen quadrant where a cross appears. Each quadrant is associated with a key on the number pad (top-left: 4, top-right: 5, bottom-left: 1, bottom-right:2). Looking instead asks that participants identify whether a target stimuli is presented on screen and if so select it's position. This increased cognitive load of remembering and identifying the target stimuli may account for some of the variation in Looking's score. When we compare Looking to the Object Judgement scores as proxies for object recognition abilities we do not get any significant correlation (Invariant: $\rho = 0.11, p = n.s.$; Identical: $\rho = 0.09, p = n.s$). This data does not indicate that Looking and Object Judgment are measuring the same underlying ability.

Looking-Flanker: Fig. B.23 compares the accuracy scores for Looking and Flanker. We see a marginal correlation in ranks ($\rho = 0.412806221, p = 0.021000981$), but an insignificant one in linear correlation ($r = 0.33077389, p = 0.069135074$).

The zoomed in data for Flanker shows a strong looking relationship between the clustered data, with noticeable outliers. When visualizing the standard deviations of the data (Fig. B.24) we get an idea of how close the outliers are to this clustered data given the wide variation in PEBL scores. Since we are working with accuracy scores between 0 and 1, small numeric variations can still create large effects on our data's correlation.

The variation could be an issue of task presentation; the items in Looking are significantly bigger and more colorful than the arrows in the Flanker test. Reduced performance could be a result of difficulty seeing the stimulus. As well, the Flanker distractors are visually closer to the target stimulus in comparison to Looking where the items are separated visually by their boxes and the player's avatar. This small variation in performance could be an issue of a couple of participants struggling early in Flanker before getting better.



(a) PEBL Flanker Scores (Mean: 0.9403, s.d = 0.0594) vs. Battery Looking Scores (Mean: 0.9397, s.d. = 0.0361), Accuracy.

(b) Zoomed in pair plots and distributions.

Figure B.23: Validity test for Looking mini-game and Flanker PEBL test.

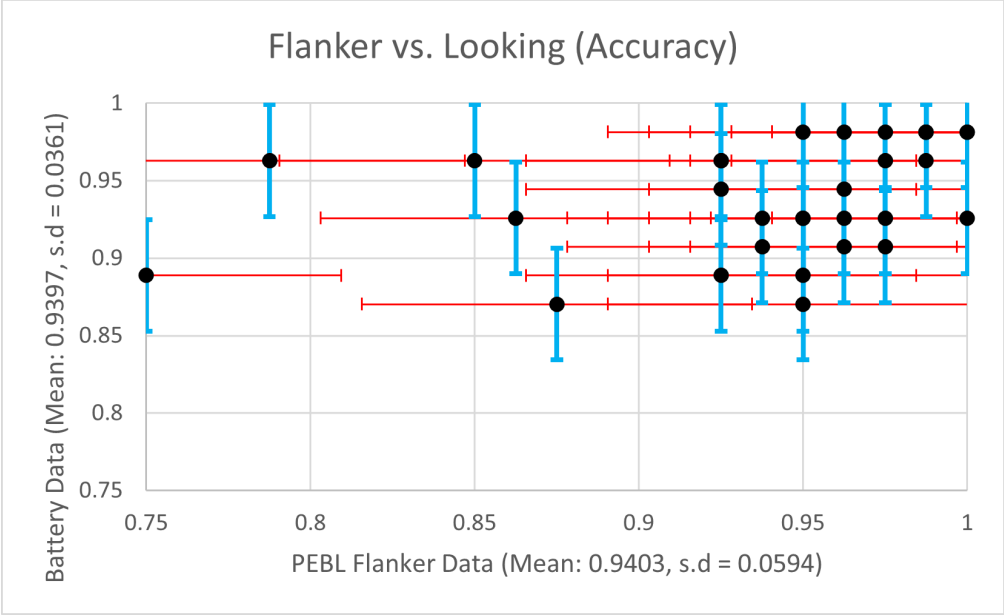


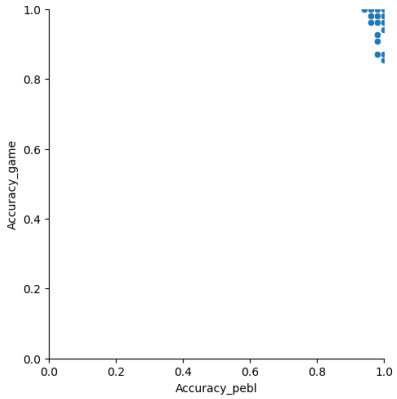
Figure B.24: Zoomed in Flanker vs. Looking accuracy data. Red bars represent the standard deviation in PEBL data. Blue bars represent the standard deviation in Looking data.

As well, Looking may be measuring multiple abilities. The marginal-but-significant rank correlation could indicate that while the abilities underlying Flanker are used in Looking, there are other unknown abilities that could be influence performance. However, with only 31 data points, we do not have enough statistical power to run a multiple regression model to see if adding other abilities would improve the relationship.

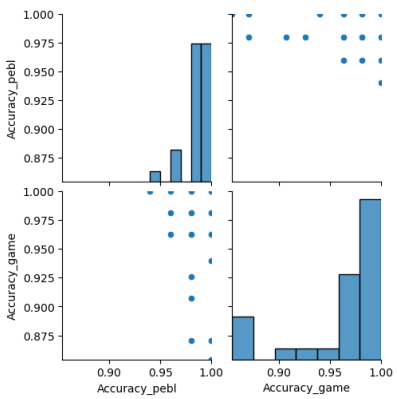
Cake-Object Judgment: Cake is the only game where it makes sense to focus on accuracy as the main measure. Therefore, we cover the details of the correlation analysis in B.4.1.

Recipe: Participant P24B only registered two instances of Recipe, and so their data was removed from the set, leaving us with 30 observations.

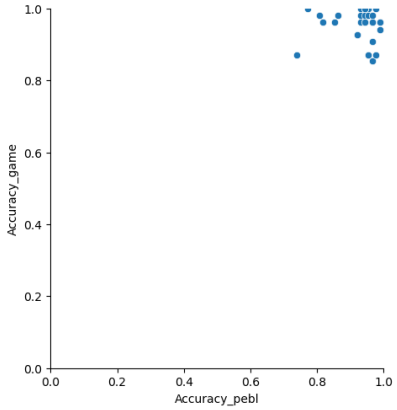
Recipe does not significantly correlate with the Four Choice Reaction Time task ($\rho = -0.3115, p = 0.093811(n.s); r = -0.29926, p = 0.108158(n.s)$), nor the Flanker Test ($\rho = -0.16135, p = 0.39433(n.s); r = 0.0239, p = 0.900239(n.s)$). Looking at the zoomed in data for both, we see Four Choice's bucketing problem. For both we see clustering but no obvious relationship to the data. This could indicate that where Four choice and Flanker measure selective attention and inhibition, Recipe measures neither of those abilities.



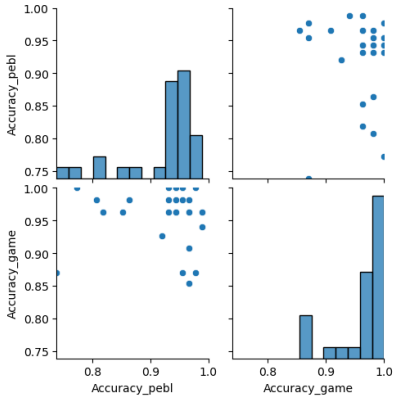
(a) PEBL Choice Reaction Scores (Mean: 0.985333, s.d. = 0.015698) vs. Battery Recipe Scores (Mean: 0.96104, s.d. = 0.04427), Accuracy.



(b) Zoomed in pair plots and distributions for Recipe and Four Choice Reaction Time Test.



(c) PEBL Flanker (Absolute) Scores (Mean: 0.925379, s.d = 0.064537) vs. Battery Recipe Scores (Mean: 0.96104, s.d. = 0.04427), Accuracy.



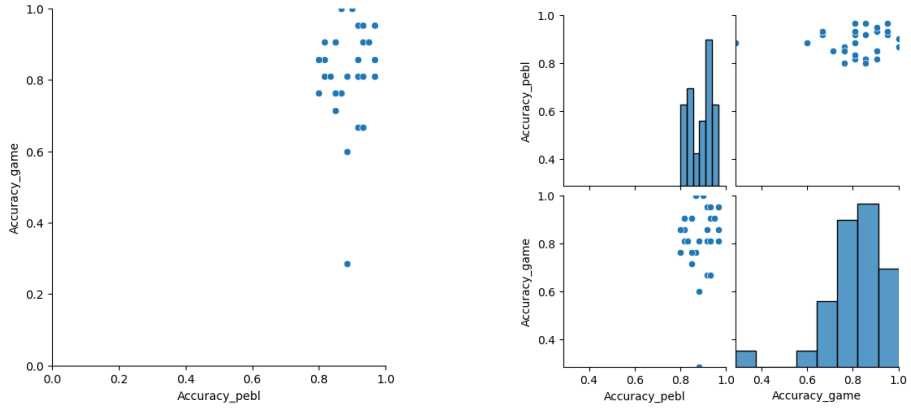
(d) Zoomed in pair plots and distributions for Recipe and Flanker.

Figure B.25: Validity test for Recipe mini-game and Four Choice Reaction Time PEBL Test.

It is possible that the design of Recipe encourages more perception and memory-based

abilities like object recognition, since participants have to recall the correct pair to pack into the box. However, when we look at the relationship between Recipe and the Object Judgment tasks, we do not see any better correlation (Invariant: $\rho = 0.00959, p > 0.9(n.s); r = -0.07366, p > 0.6(n.s)$; Identical: $\rho = 0.000313, p > 0.9(n.s); r = -0.01764, p > 0.9(n.s)$). If Recipe focuses more heavily on memory, future work could look at comparing it to an n-back test or other measures of short term memory.

Stage-Luck Vogel: Fig. B.26 confirms that there is no correlation between Stage and Luck-Vogel when looking at their accuracy ($\rho = 147370152$, $p = 0.428863793(n.s)$; $r = 0.079358321$, $p = 0.671302272(n.s)$). We can see from their distributions that PEBLs scores seem to be clustered between 0.8 and 1.0. In comparison the distribution of scores for Stages seems more reasonable. Since this is an aggregate over participant accuracies across the easy, medium, and hard conditions we take a look at the plots for each condition to see how they relate.



(a) PEBL Luck Vogel Scores (Mean: 0.8871, s.d = 0.11696) vs. Battery Stage Scores (Mean: 0.8197, s.d. = 0.1426), Accuracy. Values are averaged across the multiple trials.

(b) Zoomed in pair plots and distributions.

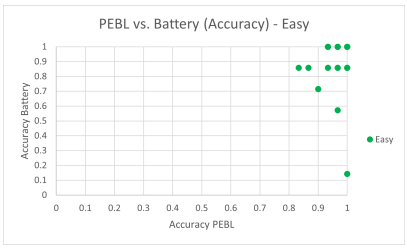
Figure B.26: Validity test for Stage mini-game and Luck Vogel PEBL test.

Given that Stage and Luck Vogel are meant to measure Token Change Detection, which is closely tied to short term memory capacity, we expected clustering in the scatter plots based on the difficulty of the task. We expected easy condition to be tightly clustered around near perfect scores, with spread increasing in both dimensions as the task became more difficult. Fig. B.27 shows the scatter plots for the different difficulties. We see none of them are correlated: Easy ($\rho = 0.125, p = 0.503(n.s)$; $r = 0.029, p = 0.878(n.s)$); Medium ($\rho = 0.137, p = 0.461(n.s)$; $r = 0.154, p = 0.407(n.s)$); Hard ($\rho = 0.109, p = 0.561(n.s)$; $r = 0.099, p = 0.598(n.s)$).

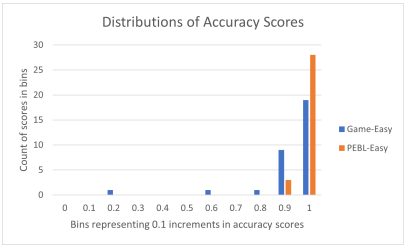
When examining the plots to see if the behaviour matches our expectations we find that the spread changes between each condition, but not as much as we would have expected. Fig. B.27b shows that PEBL had near perfect scores in the easy condition, while Stage had more

variation. The medium condition shows increased spread, but participants still performed quite well in PEBL where Stage's spread is wider (Fig. B.27d). At the Hard level where we expect most participants to do poorly, PEBL and Battery start to look more like a normal distribution (Fig. B.27f), however Stage has a surprising spike in perfect scores.

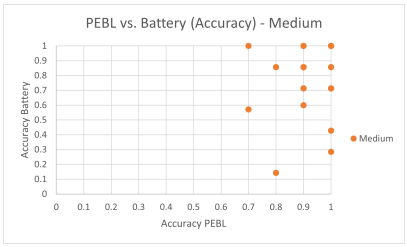
Stage's variation from PEBL at the lower condition could be because of the increased visual and audio stimuli it presents in the game-based format. As well, Stage uses the entire screen while PEBL's Luck Vogel task centralizes the stimuli so it's easier to see at once. We think the increased performance in Stage at the hard condition could be the result of increased attention. Participants made sure to tell us that the last level of Stage felt "impossible"; in observing their play participants seemed more attentive in their body language, and tried various memory techniques like speaking to themselves, and pointing at the screen. This increased attention could have resulted in higher scores. Another potential reason could be random patterns created in Stage's stimuli. As colours are randomly assigned to the on-screen characters, there were cases where colours would be unintentionally grouped. This unintentional grouping would reduce the overall difficulty of the task since the participant could more easily chunk multiple stimuli as one group instead of having to remember them separately.



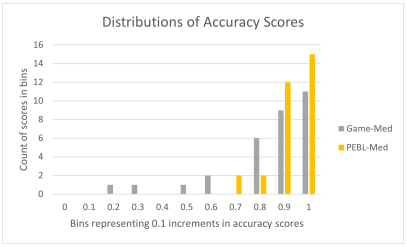
(a) PEBL Luck Vogel Scores (Mean: 0.9677, s.d = 0.1002) vs. Battery Stage Scores (Mean: 0.9078, s.d. = 0.1752), Accuracy (Easy). Values are averaged across the multiple trials.



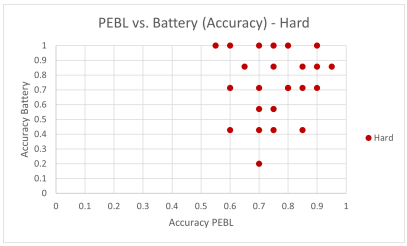
(b) Distribution of Accuracy Scores for PEBL and Battery (Easy)



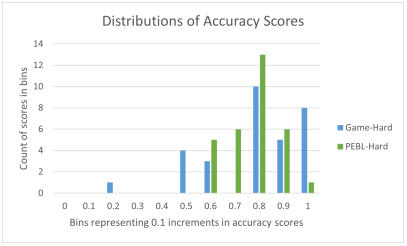
(c) PEBL Luck Vogel Scores (Mean: 0.8871, s.d = 0.11696) vs. Battery Stage Scores (Mean: 0.8197, s.d. = 0.1426), Accuracy (Medium). Values are averaged across the multiple trials.



(d) Distribution of Accuracy Scores for PEBL and Battery (Medium)



(e) PEBL Luck Vogel Scores (Mean: 0.8871, s.d = 0.11696) vs. Battery Stage Scores (Mean: 0.8197, s.d. = 0.1426), Accuracy (Hard). Values are averaged across the multiple trials.



(f) Distribution of Accuracy Scores for PEBL and Battery (Hard)

Figure B.27: Validity test for Stage mini-game and Luck Vogel PEBL test.

The way the data changes between the different test conditions shows that PEBL may be too easy to discern participant limits. This could be an artifact of the implementation of Luck Vogel for PEBL, which draws all the stimuli in a central area of the screen so participants can more easily see all the objects at once. While Stage also shows all the objects on screen at once, they are larger in the participant's field of view and competing with the background. These attributes could be making Stage significantly harder than PEBL. However, Stage's results seem to fit what we would expect to see of a sample in this

population; a distribution approaching normal, but centred around a higher performance average because the participants are more practiced at this skill.

B.3.2 Reliability

We look closer at the scatterplots for the reliability data to see any emergent trends. For reliability tests which plot battery results versus battery results, we have the Retest values on the X-axis and the Original values on the Y-axis, but this is arbitrary.

Digger. Since Digger's measure is consistent (press rate), we cover the reliability analysis in B.4.2.

Looking. P10 was removed from the dataset due to measurement errors, leaving us with 26 datapoints for this analysis. We do not see a significant correlation for Looking ($r = 0.284673025209601, p = 0.158679857351409(n.s); \rho = 0.322559413446464, p = 0.10802327549417(n.s)$). Looking at Fig. B.28a we see tight clustering in the upper right corners that could be indicative of a relationship. Zooming in with Fig. B.28b, the scatterplot appears more linear than we would expect from the correlation analysis, with one significant outlier (P25B). When considering the standard deviations of the scores, we see that the outlier is significantly further away from the rest of the data and so has a large affect on the correlation. We anticipated retest scores would vary between sessions, likely within standard deviation of the original data point. While most retest scores seem to match this, P25B is significantly outside of this range. P25B does self-identify as disabled with symptoms that can affect their attention and executive functioning (which are pivotal in selective attention and inhibition measures). It is possible that their personal variation may be larger than we anticipated since symptom flair up is inconsistent; however, we cannot make any meaningful inferences since we did not have participants rate their symptoms before sessions.

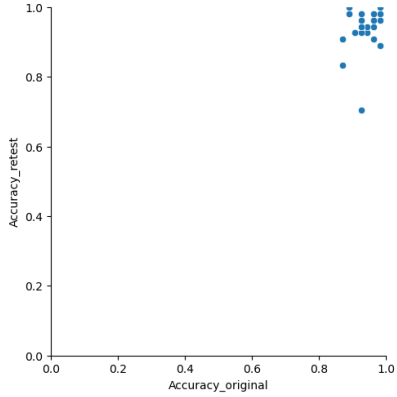
Cake. Since Cake's final measure is accuracy, we cover its reliability analysis in B.4.2.

Recipe. There is no significant correlation for Recipe ($r = 0.183565239160104, p = 0.359393883829444(n.$ $0.24867721662188, p = 0.211013936960891(n.s)$). Fig. B.29 seems tightly grouped; the closer view shows some outlier and evenly spaced differences between the clustered data in the top right. It is unclear why the retest reliability is poor here.

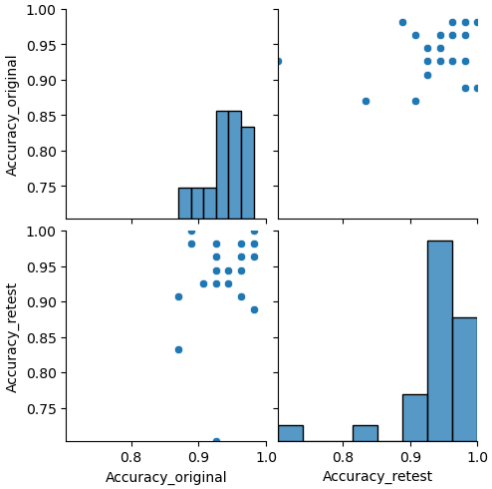
Stage. There is no significant correlation for Stage ($r = 0.150349p = 0.44507(n.s); \rho = 0.136582, p = 0.488289(n.s)$). Looking at the average data in Fig. B.30 we see no obvious pattern in the data. However, we do see performance in the retest is more normally distributed.

B.4 Detailed Data Analysis - Final Measures

This section covers the detailed analysis and discussion of the correlation study data. We assess both the validity and reliability of the games using their appropriate measures.

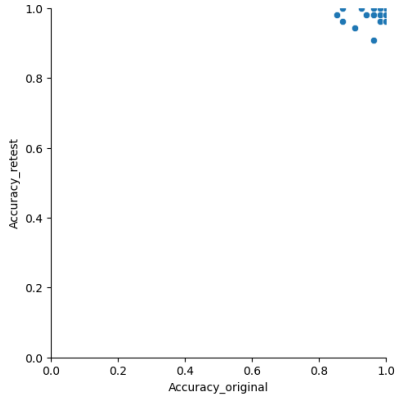


(a) Retest vs Original Looking scores, Accuracy.

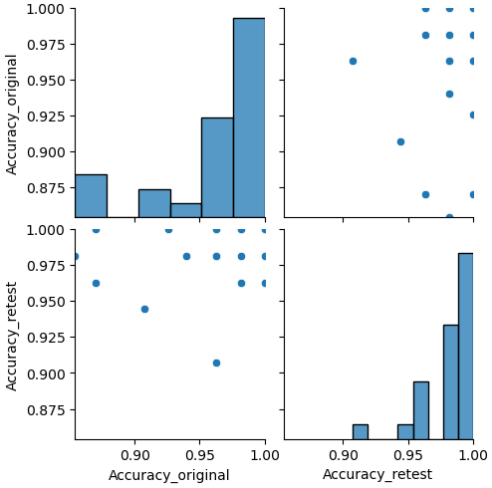


(b) Zoomed in Looking reliability data. Red bars represent the standard deviation in Retest data. Blue bars represent the standard deviation in Original data.

Figure B.28: Reliability test for Looking mini-game.



(a) Retest vs Original Recipe scores, Accuracy. Values are averaged across the multiple trials.



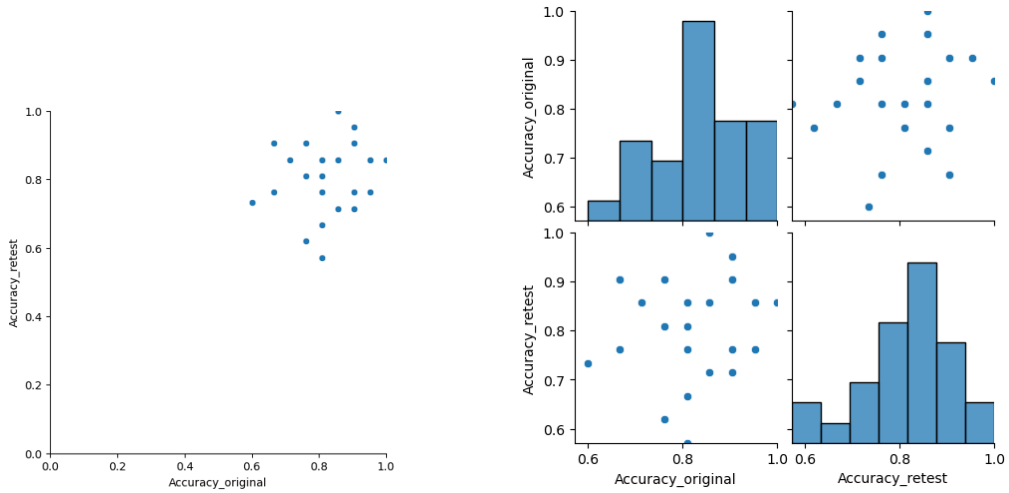
(b) Zoomed in Recipe reliability data.

Figure B.29: Reliability testing for Recipe.

B.4.1 Validity

We look at each pair of mini-game and PEBL test based on the appropriate measure: press rate (number presses per second), accuracy (mean correct responses), or rate correct score (RCS: correct responses per second). Wherever scatter plots are presented we use PEBL scores as the independent variable (X-axis) and Battery scores as the dependent variable (Y-axis) to simplify reading and comparison.

Recall in reading this section we are testing H1:



(a) Retest vs Original Stage scores, Accuracy.

(b) Zoomed in Stage (Avg.) reliability data.

Figure B.30: Reliability testing for Stage (Avg.).

H1 For the mini-game battery to be considered valid participants' mini-game assessment scores must positively correlate to the associated existing measurement scores, with $\rho \geq 0.7$ or $r^2 > 0.5$.

We set $\alpha = 0.05$, and consider any correlation to be significant if the p-value $< \alpha$.

Digger-Tapping: Digger and Tapping are compared via press rate. There was a measurement error with P03's PEBL data, so it was removed from the set. We still have 30 participants in this pool so the data is viable. The accuracy data seems relatively correlated (Fig. B.31), with both Spearman ($\rho = 0.770857, p < 0.0001$), and Pearson ($r = 0.741397, p < 0.0001$) correlations showing significance and strength.

Looking-Four Choice: Looking and Four Choice are compared using RCS. No data was removed from this set in cleaning. We see both the Spearman ($\rho = 0.375, p = 0.03765$) and Pearson ($r = 0.3728, p = 0.03885$) correlations are significant. Fig. B.32a shows an obvious pattern between these two measures. We see that the values for Looking's RCS exist in a very narrow range; this is confirmed when looking at the distribution in Fig. B.32b. This is potentially because Looking was more difficult than Four Choice in its design.

Unlike Four Choice, Looking actually requires players to choose between five option (the four directions, and the "not shown" input). While this means that players may spend more time searching for the stimuli before responding, the bigger impact is participant's forgetting that there was a "not shown" option. We had multiple participants ask us if the game was broken or what they should press when they first encountered the "not shown" conditions. While this could be a situation where participants did not clearly read the instruction, it could also be the case that participants became engrossed in the selection response and so the task switching to the "not shown" condition was so high they "forgot".

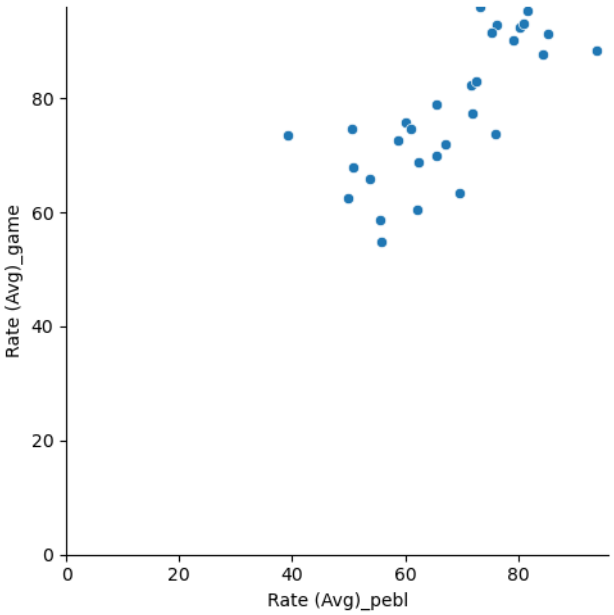
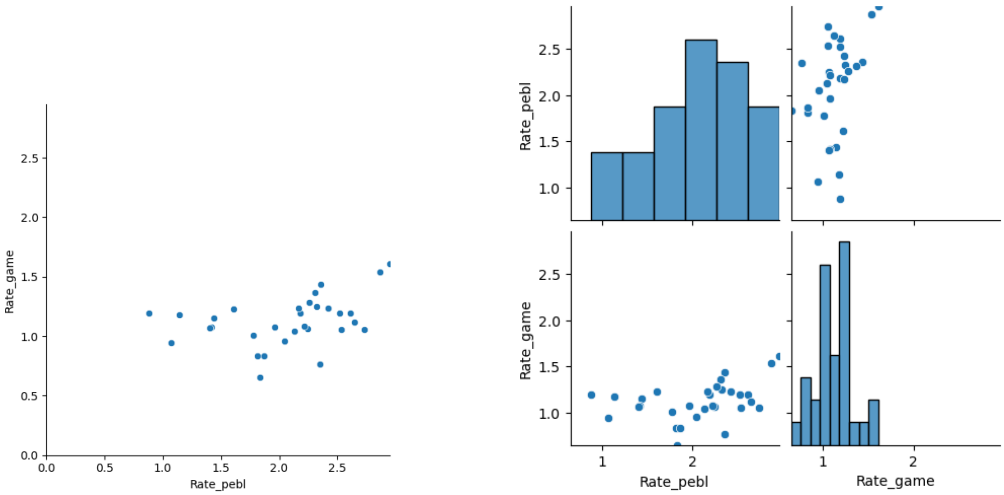


Figure B.31: Validity test: PEBL Tapping Scores vs. Battery Digger Scores, measured in Presses/Second.



(a) Four Choice vs Looking scores, Rate Correct Score.

(b) Zoomed in pair plots and distributions for Four Choice and Looking scores.

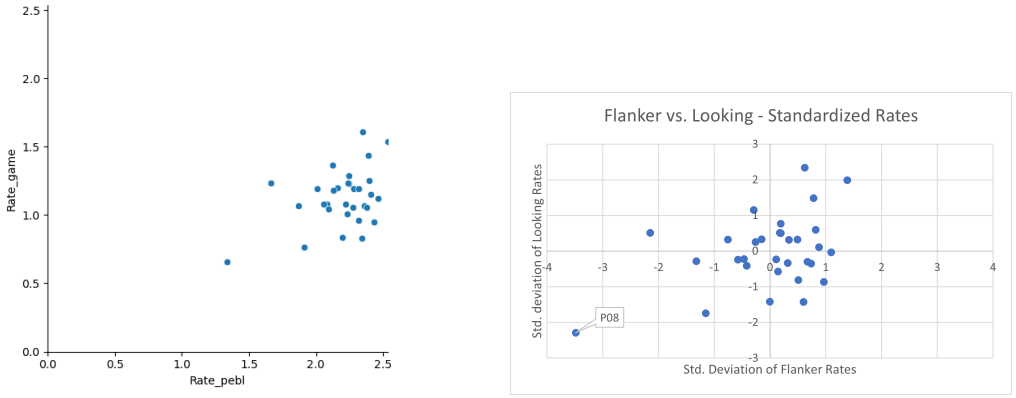
Figure B.32: Validity correlation results for Looking and Four Choice.

Another potential reason for this restricted range is Looking's distractor stimuli; the item set for the game featured similar items to increase the difficulty of selecting the correct object. We had multiple participants mention to us after their sessions that the distractor objects made the game "hard" because they kept picking them while trying to move quickly. This was the purpose of the distractor objects, as avoiding them would require that participants inhibit their response.

Another reason this range may be limited is the way that Looking was presented to the players at increasing difficulty levels. Difficulty was controlled by the number of stimuli changes in an instance; the easy level would not change the target stimuli during the trials, while the hardest level changed the stimuli every three trials. Participants noted that the "stimuli may change" wording was confusing in the instructions as many were waiting for the stimuli to change *every* trial, and so wasted time in the easy condition figuring out that the stimuli would not change.

With all of these reasons for the data to look like it does, we wonder whether Looking, by design, incorporates abilities other than selective attention and inhibition. The distractor items seem to create some object recognition situations, as perceived by the participants. However, when compared against the object recognition PEBL tasks we do not see any significant correlation to indicate this might be the case.

Looking-Flanker: Looking and Flanker are compared via RCS and use all 31 datapoints. Our original calculations of Spearman ($\rho = 0.210, p = 0.255724$) and Pearson ($r = 0.424, p = 0.017475$) are at odds. Looking at the scatter plot (Fig. B.33a) we see there is an outlier in the data. Fig. B.33b shows that P08's data is over 3 standard deviations under the mean for PEBL data so we consider removing it from the set as it is exerting a significant amount of influence on Pearson's calculation.



(a) Flanker vs Looking scores, Rate Correct Score.

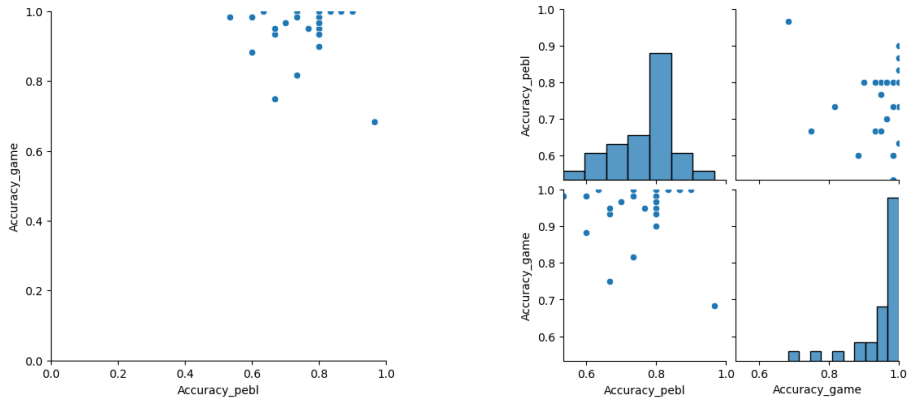
(b) Standardized rates for Flanker and Looking, to show deviations from mean.

Figure B.33: Validity correlation results for Looking and Flanker.

Looking into P08 we see this participant self-report playing predominantly Clash Royale on their smartphone/tablet. They report a lower than average amount of gaming frequency (once a week) and time (15 to 30 minutes) compared to the average participant. The type of deck-building tower defense gameplay they prefer does not require the same intensity of selective attention/inhibition proficiency as other game types, which could explain poorer performance in inhibition tests like Flanker. As well, our observational notes list that they generally seemed tired and frustrated with the PEBL tasks from their body language and comments. This tiredness and frustration outside of the control of our experiment could be a reason why they performed so out of line with the other participants. Given that we have reason to believe their performance is impacted by external factors (tiredness, disengagement

with testing method) we believe it is reasonable to remove P08 from the dataset. Removing this point, Spearman ($\rho = 0.129, p = 0.498$) and Pearson ($r = 0.216, p = 0.251$) are in alignment about the data being insignificant.

Cake-Object Judgment: We see a marginal rank correlation between Cake and the Invariant condition ($\rho = 0.362135846, p = 0.045289$) but a significantly different Pearson correlation ($r = -0.0215, p = 0.908649(n.s)$). There is no significant correlation between Cake and the Absolute condition ($\rho = -0.0075, p = 0.968057(n.s); r = -0.08596, p = 0.645649(n.s)$). In some ways this makes sense, as Cake is closer in task design to Invariant than Absolute. For Cake participants must sort familiar items into their respective categories. Invariant tasks participants to select which object matches the one they previously saw (irrespective of scale and rotation). Absolute asks player to select the object that perfectly matches the previous object in shape, scale, and rotation. While they are similar in design, this does not explain the difference between the Spearman and Pearson correlations, nor does it explain why the Spearman relationship is fairly weak.



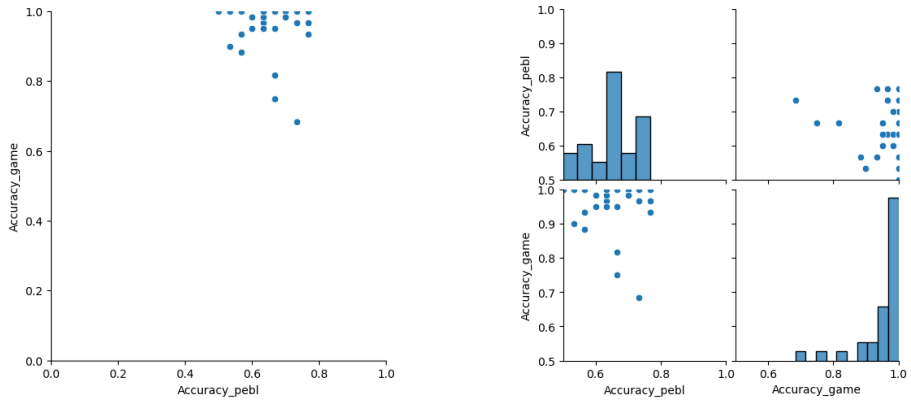
(a) PEBL Object Judgment (Invariant) Scores (Mean: 0.7645, s.d = 0.0954) vs. Battery Cake Scores (Mean: 0.9543, s.d. = 0.0766), Accuracy.

(b) Zoomed in pair plots and distributions for Invariant.

Figure B.34: Validity tests for Cake mini-game and Object Judgments Invariant PEBL tests.

Looking at the scatterplots for the Invariant (Fig. B.34a) and Absolute (Fig. B.35a) conditions shows some clumping, but many outlier points. The zoomed images for the data shows many participants ended up with a perfect score on Cake, while PEBL seems to have better spreads. This leads us to believe that Cake was too easy. Thinking about its design, the difficulty should be coming from how quickly and accurately the participant can identify the object. Cake asks players to identify the object quickly enough to move the sorter to the correct category. Since the items move down the conveyor belt at a constant speed, difficulty should be based on how quickly the participant could recognize and categorize the item. Since the items are visually distinct (e.g. pizza, cupcake) and have obvious categories (e.g. entrée, dessert), the categorization may have been too easy. The difficulty is modulated by how fast items spawn on the conveyor belt which acts as both a distractor for the current item to be sorted, and as a way to reduce the amount of reaction time the participant has

between objects. While the distinct and familiar objects were selected to make the gameplay accessible, this may be a case where the items and categories should be custom to the gameplay in order to have players enact their ability more.



(a) PEBL Object Judgment (Absolute) Scores (Mean: 0.6527, s.d = 0.0749) vs. Battery Cake Scores (Mean: 0.9543, s.d. = 0.0766), Accuracy.

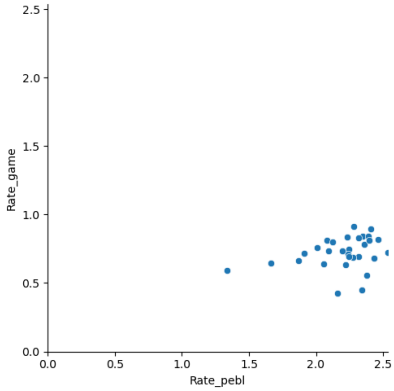
(b) Zoomed in pair plots and distributions for Absolute.

Figure B.35: Validity tests for Cake mini-game and Object Judgments Absolute PEBL tests.

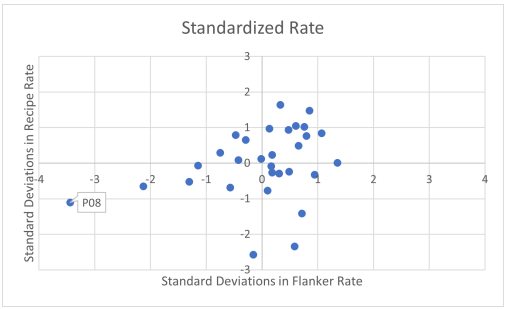
Recipe-Flanker: Recipe and Flanker are compared along RCS. We removed P24 from the dataset due to measurement errors, meaning we are analysing 30 data points. Recipe is not significant for Flanker (Spearman $\rho = 0.340600667, p = 0.065507$; Pearson $r = 0.294709038, p = 0.113891$). As we saw in Looking, P08 for Flanker is more than 3 standard deviations below the mean, and we have reason to believe the impact on their score comes from external factors not individual differences, so we remove it as an outlier. This brings us under our minimum 30 data points. While not able to make any claims with them, we calculate the correlations to see how they look (Spearman $\rho = 0.296, p = 0.120$; Pearson $r = 0.214, p = 0.265$).

Recipe-Four Choice: Recipe and Four Choice are compared along RCS. Recipe removes P24 from the dataset due to measurement errors; this leaves us 30 data points for our analysis. Recipe is significant across both Spearman ($\rho = 0.402002, p = 0.027655$) and Pearson ($r = 0.435002, p = 0.016289$) coefficients. Fig. B.37a shows data with a similar pattern to Looking-Four Choice, with the mini-game rates bounded. When looking at the distributions in Fig. B.37b we see that the distribution seem normal, though Recipe's is heavily compacted.

Unlike Looking, Recipe is simpler than Four Choice; participants must just decide whether the presented candies matches the target set (i.e. a Go/No-go task). While cognitively simpler, Recipe employs distractor candy that may be increasing its difficulty. In both Recipe and Looking the inclusion of the distractor items makes accuracy more rewarded than speed, where four choice seems to cause more prioritization of speed from participants.

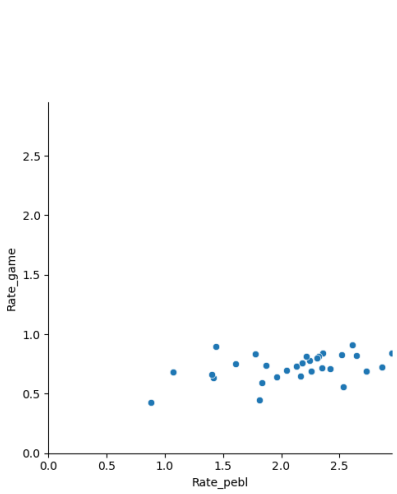


(a) Flanker vs Recipe scores, Rate Correct Score.

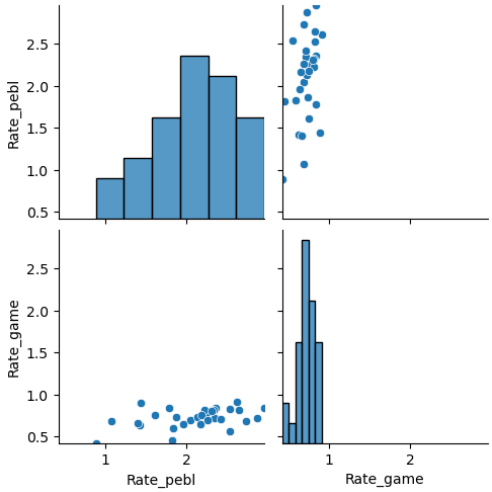


(b) Standardized rates for Flanker and Recipe, to show deviations from mean.

Figure B.36: Validity correlation results for Recipe and Flanker.



(a) FourChoice vs Recipe scores, Rate Correct Score.



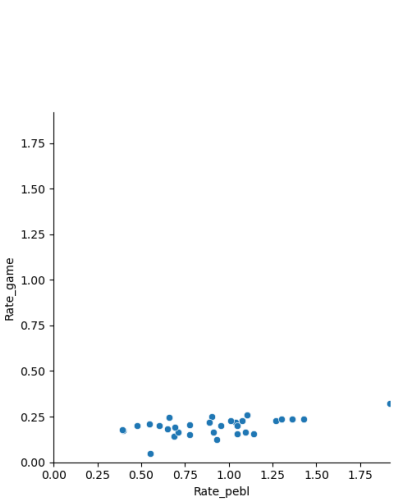
(b) Zoomed in pair plots and distributions for Four Choice and Recipe.

Figure B.37: Validity correlation results for Recipe and FourChoice.

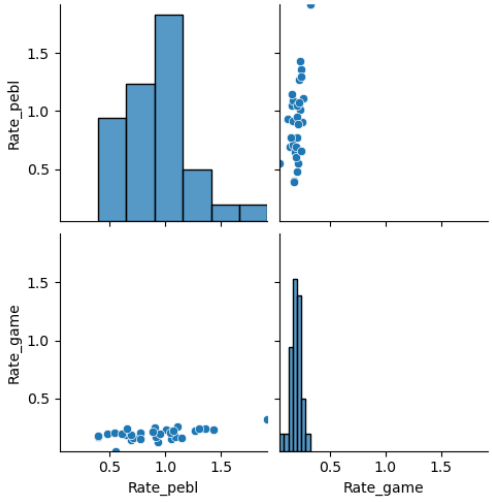
Like Looking, participants were also unsure of how frequently the target stimuli would change. In the easy condition of Recipe, the stimuli never changes. However, participants who skimmed the instructions or did not understand that "may change" meant it could stay the same, paused during their trial because they were waiting for a new target to show on screen. This waiting time in the early conditions, accompanied by the difficulty of distractor items may be why rates are seemingly concentrated between 0.5 and 1.

Stage-Luck Vogel: Stage and Luck Vogel are compared via RCS, and use all 31 data points. When we originally examine them, we get very strong and significant correlations for both Spearman ($\rho = 0.443548387, p = 0.012444$) and Pearson ($r = 0.555854862, p = 0.001168$). We see this correlation in Fig. B.38a; its shape and pattern seem similar to

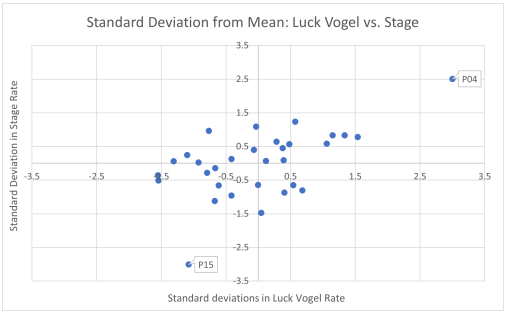
Looking-Four Choice and Recipe-Four Choice. However, we notice at the far right of the axis a datapoint which looks like it may be significantly affecting our results. Taking a look at the standard deviations (Fig. B.38c), we see two datapoints that are significant outliers (P04 significant Luck Vogel outlier; P15 significant Stage outlier). Removing these two data points we no longer have sufficient power to draw conclusions about the correlations (though we calculate them as Spearman $\rho = 0.345, p = 0.067$, and Pearson $r = 0.350, p = 0.063$).



(a) Luck Vogel vs Stage scores, Rate Correct Score.



(b) Zoomed in pair plots and distributions for Luck Vogel and Stage.



(c) Standard deviation for Luck Vogel and Stage rates.

Figure B.38: Validity correlation results for Stage and Luck Vogel.

Looking at the participants attached to these data points, it makes sense that P04 performs at such a high level considering they are a self-reported gamer playing a variety of games 5 or more times a week. The games listed by P04 (Fortnite, Super Smash Bros.) require significant attentional control and so may indicate significant practice with token change detection skills that explain the high score. In comparison P15 reports that they used to be a gamer but do not play enough to consider themselves one anymore; they list playing 2 to 4 times a week but only list playing Rocket League. It is unclear from the score why they may be performing poorly given this self-reported gaming history. P15 was recruited in the final set of participants, and we have noted in our session notes that they had just finished their exams and were completing this experiment before going home for

the break. It could be that P15 was mentally fatigued from circumstances outside of this experiment.

Given the individual context of P04, this deviation from the mean seems reasonable and reflective of individual differences in our participants leading us to consider including it in the data. In comparison, P15 does not have a strong reason why it should be included in the data, but the external factors (mental fatigue) are assumed from their comments about their life not observations of body language or specific comments being frustrated or tired. If we include P04 but drop P15, we would have our sufficient minimum and the rates become strongly and significantly associated again (Spearman $\rho = 0.409, p = 0.025$, and Pearson $r = 0.546, p = 0.002$). This implies that P04 has a stronger effect on the correlation (likely because of how far from the mean it falls). Since the more impactful data point is being kept, and there is not a clear enough reason to remove the other data point we will keep both in the data set and stick with the original correlation analyses.

B.4.2 Reliability

We look at each mini-game test-retest pair based on the appropriate measure: press rate (number presses per second), accuracy (mean correct responses), or rate correct score (RCS: correct responses per second). Wherever scatter plots are presented we use original test scores as the independent variable (X-axis) and retest scores as the dependent variable (Y-axis) to simplify reading and comparison.

Recall in reading this section that we are testing H2:

H2 For the mini-game battery to be considered reliable participants' mini-game assessment scores between sessions must positively correlate with a $p \geq 0.7$.

We are assessing reliability using Pearson correlation; we also calculate the Spearman correlation and intraclass correlation coefficient (ICC) to get a more robust understanding of the mini-game's reliability. ICCs are calculated in SPSS as an absolute agreement coefficient using a single measure, two-way mixed effects paradigm. While it is imperative to use an absolute agreement ICC for test-retest reliability [240], we recognize that people will perform differently during different sessions due to a variety of factors. As such we do not expect perfect agreement in the scores, and do not believe it to be a reasonable goal for our work.

Digger. Digger only uses 27 data points, since P03's data was not captured properly in the original test. Digger shows strong correlations between the test-retest press rate data (Fig. B.39), with both Pearson ($r = 0.825, p < 0.0001$) and Spearman ($\rho = 0.811966, p < 0.0001$). ICC reports correlation of $0.793, p < 0.001$. This leads us to believe the reliability for Digger is good.

Looking: Looking uses 26 data points as P10 needed to be dropped due to measurement errors in the retest. The given data (Fig. B.40) shows strong correlations between the test-retest data for Pearson ($r = 0.657047, p = 0.000266$) and Spearman ($\rho = 0.603419, p = 0.001101$). ICC reports a correlation of $0.577, p < 0.001$. This leads us to believe that the reliability for Looking is moderate.

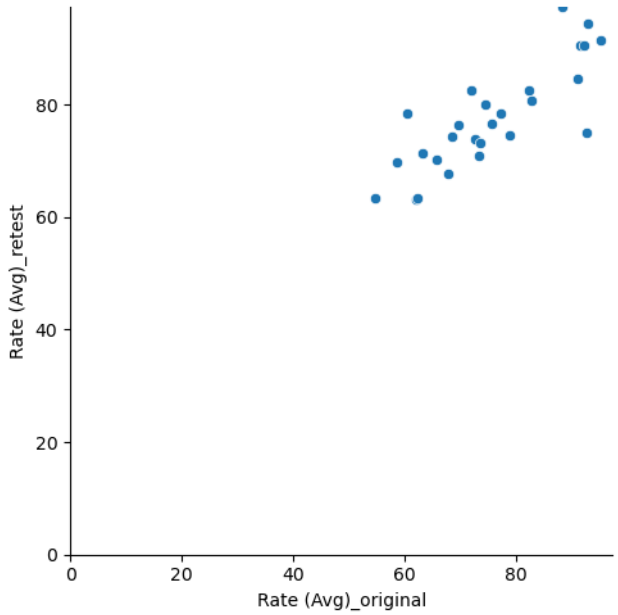
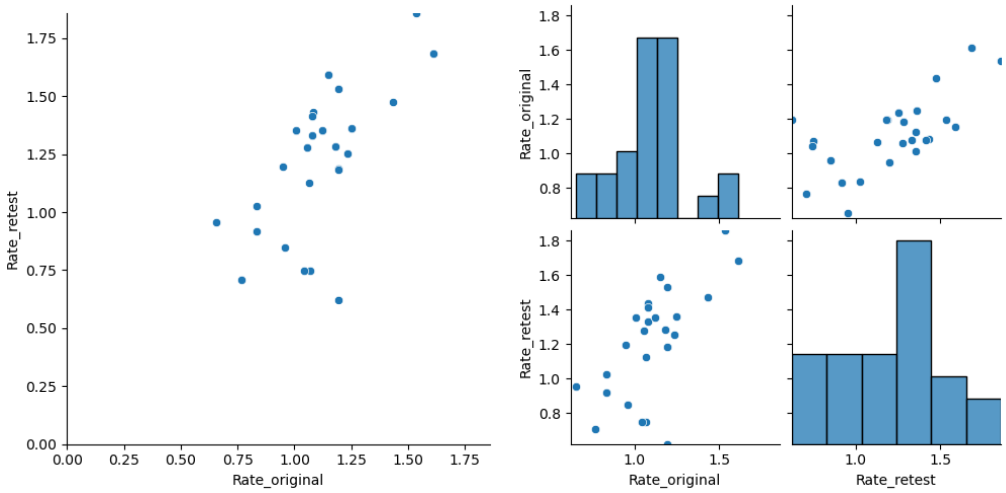


Figure B.39: Reliability test: Retest vs Original Digger scores, measured in Presses/Second. Values are averaged across the multiple trials.

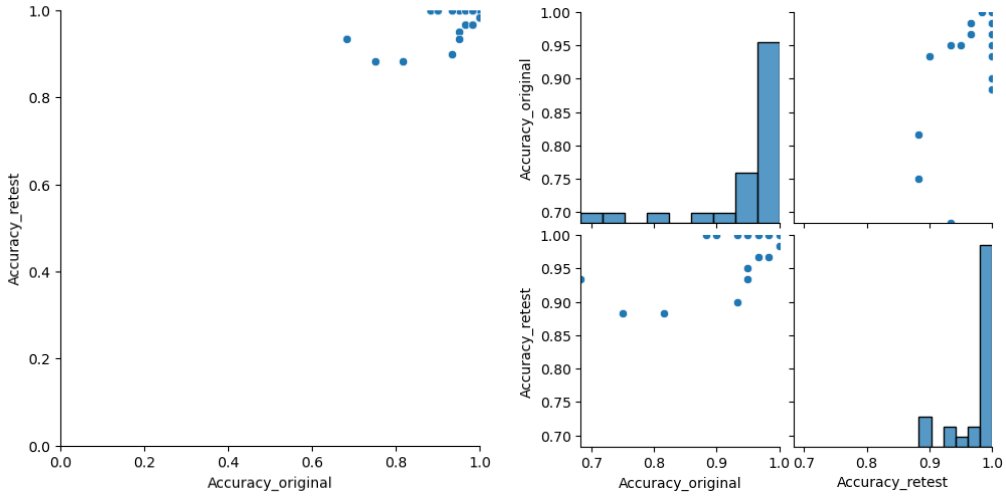


(a) Retest vs Original Looking scores, (b) Zoomed in Looking reliability data. Rate correct score.

Figure B.40: Reliability testing for Cake.

Cake. Cake scores are somewhat reliable (Pearson: $r = 0.6818, p < 0.0001$; Spearman: $\rho = 0.4863, p = 0.008695$). ICC indicates poor reliability (ICC: $0.486, p = 0.002$), but given how close it is the 0.5 "moderate" threshold and taken in conjunction with the other correlation measures, we believe Cake can be thought of as moderately reliable. There are a couple odd outliers in the data (Fig. B.41) which show that participants did better in the retest than in the original. With Cake being too easy, the lower scores on the original test could be due to testing fatigue or experimental errors like participants getting a little bored

and not paying attention.



(a) Retest vs Original Cake scores, Accuracy. Values are averaged across the multiple trials.

(b) Zoomed in Cake reliability data.

Figure B.41: Reliability testing for Cake.

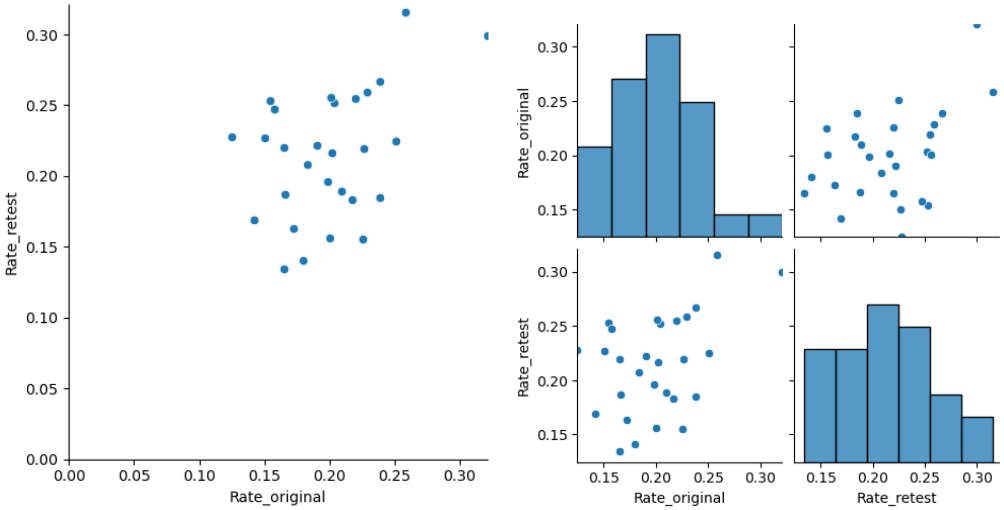
Recipe: Recipe uses 27 data points as P24 is dropped from the original tests due to a measurement error. Data shows a moderate correlation with Pearson ($r = 0.597433, p = 0.001$), and Spearman ($\rho = 0.680708, p < 0.0001$), but an unreliable ICC ($0.388, p < 0.001$). This leads us to believe that the reliability for Recipe is somewhat reliable, though Looking seems to be better across Pearson and ICC. This is something to consider as both games target the same abilities, and correlate to the same PEBL tasks.

Stage: Stage uses 28 data points. Stage exhibits weaker correlations than the other games, with Pearson ($r = 0.424584, p = 0.024321$), Spearman ($\rho = 0.316913, p = 0.10035$), and ICC ($0.405, p = 0.011$). Taking a look at the reliability data scatterplots in Fig. B.42, we see no clear pattern emerge from the data.

It is unclear why Stage performance does not show a clear relationship. Some possibilities are practice effects from the original session affecting the retest measures, reduced fatigue in the retest leading to higher scores, or lucky trial set ups in one or the other.

When participants return for the second session, it was assumed there was enough time between session that participants could be considered "fresh" when interacting with the mini-games. What we noticed instead was that participants remembered their perceptions of certain games being hard or easy. Since most participants had trouble with Stage in the original session, we observed that some were more attentive and focused during the Stage instances in the retest. This could have led to improved performance.

Another factor could be that the retest is significantly shorter than the original, and so participants may be less fatigued by the time they reach Stage. While fatigue effects were mitigated in the original session by randomizing participant task order, for reliability we only



(a) Retest vs Original Stage scores, Rate correct score.

(b) Zoomed in Stage reliability data.

Figure B.42: Reliability testing for Stage.

care about one game at a time so these effects may come through in the data. We partitioned the data to compare Group A (PEBL-first) and Group B (Battery-first) participants; each group had 14 participants. Group A showed stronger reliability correlations than Group B. For Group A, we found Pearson ($r = 0.623, p = 0.017$), Spearman ($\rho = 0.547, p = 0.043$), and ICC ($0.615, p = 0.007$) which indicates a moderate reliability for this group. For Group B, we found Pearson ($r = 0.131, p = 0.656$), Spearman ($\rho = 0.103, p = 0.725$), and ICC ($0.131, p = 0.321$) which indicates poor reliability for this group. Given this information, we see there is some kind of effect happening between groups but do not have enough information to determine exactly what this effect is.

The final aspect we think may affect our reliability measures for Stage is the randomized trial design. Since the colour of the stimuli are randomized every trial, it is possible that participants encounter trials that are significantly easier that the other trials in the set because the randomized colours unintentionally create patterns. These patterns allow for easier chunking of information in memory, which means that participants may perform significantly better in these cases than we would expect them to. It is possible that these randomized "lucky" trials may have influenced the test/retest scores depending on if and when they were encountered.

Appendix C

Timing Challenges

In Ch. 14 we noted that two mini-games did not meet the button mashing family when we re-examined them. These two challenges fit our view of Timing challenges. Given this we present here a preliminary case study of Ballon Burst as a timing challenge.

C.1 Gameplay Case Studies

Balloon Burst [193, 194, 323] has players inflating Bowser-shaped balloons using a hand pump. In Mario Party players are doing so individually, while in Mario Party 2 and Mario Party Superstars players work in teams of two. For the purposes of this analysis we will be discussing the Mario Party 2 version played against (and with) CPUs (Fig. C.43). The gameplay lasts for 30 seconds, with the timer positioned in the top middle of the screen. Players alternate pressing the A and Z/B buttons (R and L in Superstars) respectively to push the pump down and pull it up, thus filling the balloon with air. The winning team is the one who finishes filling their balloon first. Players are advised to press the next button in sequence when the pump flashes, as that indicates it is "full of air". In this way the game controls the pacing of button mashing, and signals what part of the sequence the player is on. The context of the game is local team-based competitive multiplayer on a standard controller.

The main motor interactions for this game are finger pressing and wrist pointing. Fingers are used to interact with the buttons, while wrist pointing stabilizes the wrist and hand, thereby helping with the finger pressing speed. Speed of pressing these buttons is fairly important in this game, as the faster these buttons get pressed, the faster the balloon inflates.

While speed is important, the effectiveness of just button mashing is low for most levels of difficulty. In this way, we see the advice of waiting for the pump to flash gives us deeper insight into *how* the game is designed to be played. As well, the actions mapped to the button presses account for how long the button is being held. The A button needs to be held until the pump is fully depressed, and then the Z/B button needs to be held until the pump is fully inflated (and starts flashing). Rather than speed, an understanding of timing is therefore necessary to be competitive.

Inhibition, token change detection, and selective attention are the three abilities that inherently address the timing aspect this mini-game. Inhibition is used to refrain from

Figure C.43: Video clip of Mario Party 2's Balloon Burst gameplay from Youtuber NintendoMovies. Click image to play.

button mashing, and keeps us waiting for the "right" time to act. This is important as the other factors of the game (competitive multiplayer, time limit) encourage the player to rashly button mash. Token change detection helps us recognize the state of the pump as the visual stimulus changes. When we recognize the pump as deflated we know to press the Z/B button; when we see the pump flash we know to switch to pressing the A button. Token change detection is the ability which detects the trigger we are inhibiting behaviour around. The faster we can detect the change, the quicker we can respond to it. Selective attention ties these together; it keeps us focused on the gameplay, which is longer than most other mini-games, and tells us to act once we are triggered by the pump flashing. Selective attention also works with inhibition to keep us from attending to other distracting/irrelevant visual stimuli like the other pumps on screen, the pressure gauges, etc. Without these three abilities working in tandem, pressing the buttons in the correct order is irrelevant as it doesn't optimally inflate the balloon.

Tactile perception and procedural memory are abilities that are used but significantly less important to the gameplay. Procedural memory is used in remembering the layout of the controller and what buttons to press. Considering the length of the sequence is short, and there are multiple options for pressing (Z or B), this is minimal in effort. Tactile perception supports procedural memory as the player can feel the buttons (which on the N64 Controller are distinct in size and position). This means which button is being pressed, and whether it is pressed or not is something we can feel/perceive quickly and easily.

Considering the gameplay, we believe that there are two abilities which bottleneck play: token change detection and finger pressing. This reflects the importance of both reacting and speed to successful gameplay completion. The gameplay is inherently a race, implying finger pressing speed as the limiting ability. However, winning this race is contingent on well timed presses triggered by quick change detection. We can't quantify the difference between them easily, but since pure speed can't win this challenge alone we decide to place token change detection as marginally more important than finger pressing, even though both are limiting

abilities.

Since the timing element of the race is inherently important, we believe inhibition and selective attention are important but not limiting abilities. This is because they support the reaction to the changing pump. Tactile perception and procedural memory are used, but not noticeably. We believe tactile perception is slightly more important than procedural memory because of the feedback it gives us about the button presses. Since the buttons need to be held for a short time, not just pressed and released quickly, knowing that it is still depressed and when we have released it is useful in our timing.

With all of this in mind, we consider the priority ordering of abilities for this gameplay to be:

1. Token change detection
2. Finger pressing
3. Inhibition
4. Selective attention
5. Tactile perception
6. Procedural memory

Appendix D

Custom Button Mashing Challenge Designs

We identified three types of button mashing challenges: single input, multiple input, and alternating input. We aim to test the competency profiles for all of these challenges. Thus we need three separate customizable games.

The games must fit the definition, mechanics, and context for each button mashing challenge. The main caveat is that our games are built to be played with the keyboard, not standard controllers. Since we aim to use these challenges in this and potentially future controlled studies, it is important that the variable aspects which affect the nature and difficulty of the game are parameterized. This way we can easily change elements to test their effects on perceived and measured difficulty. The games must also output participant score information for easy analysis. Finally, we want these games to look and feel cohesive to promote that "game" feel.

A Quick Aside...

Given that we are looking to study performance against measured abilities, you may wonder why the mini-games give the players ranks at all. Our original implementation provided the same animation for all players regardless of performance and told them they did a "good job". During our internal testing of the game, players said they realised in the practice that "it didn't matter how they performed" and so they did not try as hard as they could have. We added the different ranks and animations to incentivize engaged performance as we care about players actually trying. This aligns with trying to promote "game" feel.

To this end we created three games: Fire Starter, Fly Away, and Potion Master. These three games place the player as a young witch practicing her magic skills. All challenges were developed in Unity 2022.3.0f1 using the C# scripting language. They all output the game name, default parameters, and player's score. The games use free assets from various locations, the details of which are documented in the credits of the game. The source code for the games, along with an executable, are available on GitHub.

D.0.1 Single Input Button Mashing: Fire Starter

Figure D.44: Single input button mashing challenge designed to be customizable.

Fire Starter (Fig. D.44) tasks the player with rapidly and repeatedly hitting the specified button within the allotted time span to light a fire. The parameters for the game, along with their default values are found in Table D.2.

Parameter	**Default Value**	**Explanation**
Button	Space bar	The specific button pressed may affect difficulty due to its size and postion. We default to space bar as it is large and central, making it unlikely for participants to "slip" and miss.
Time limit	10 seconds	The time given to the player limits how many times they can press the button. 10 seconds is the default as it reflects the way average tapping speed was measured (60 presses/10 seconds), and is similar in length to many classic button mashing Mario Party minigames.
Goal	60	The expected number of presses is the main measure of difficulty for SIBM as this is the "success target". It defaults to 60 presses as that was the measured "normative" data.
Score modifier	1	This affects how a button press is counted towards the goal. It provides a way to artificially induce fatigue or skew the difficulty. 1 is the default because it is the intuitive way players will understand their score increasing (1 press = 1 press).

Table D.2: Parameters of SIBM custom game — Fire Starter — with the listed default settings and an explanation of the parameter's purpose.

The player's score is calculated as # button presses × score modifier. For a player in the default configuration to get a score of 60 (i.e. the goal) they need to press the space bar (correct button) 60 times. This default was chosen based on previous literature on average finger pressing rates. The final score will determine whether the fire is lit, and what colour it burns as per Table D.3. These values are fairly arbitrary as they are meant to provide

feedback to the player and encourage their play. Since the game outputs final scores for our analysis we do not need to be specific about this subdivision.

Rank	Player final score	Fire State	Fire Colour	Effect Meaning
F	$< 0.5*$Goal	Not lit	–	Really bad
D	$0.5*$Goal $\geq$ Score $< 0.75*$Goal	Lit	Green	Bad
C	$0.75*$Goal $\geq$ Score $<$ Goal	Lit	Yellow	Low Average
B	Goal $\geq$ Score $< 1.25*$Goal	Lit	Orange	High Average
A	$1.25*$Goal $\geq$ Score $< 1.5*$Goal	Lit	Blue	Good
S	$\geq 1.5*$Goal	Lit	White	Really good

Table D.3: List of fire states and colours based on player final score relative to the goal number.

D.0.2 Multiple Input Button Mashing: Fly Away

Figure D.45: Multiple input button mashing challenge designed to be customizable.

Fly Away (Fig. D.45) shows our witch preparing to take off in flight. The player charges her spell by rapidly pressing two target buttons simultaneously within the allotted time. The parameters for the game, along with their default values are found in Table D.4. Since we were unable to find literature about simultaneous button pressing rates, this goal is based on the results of the pilot testing for the games where players performed relatively similarly to their single input button mashing scores thus making the "same" goal seem reasonable. This makes sense if players effectively approach the two games the same way: if pressing a single button means locking your finger in position and pivoting at your wrist, then pressing two buttons means locking two fingers but the wrist movement and speed is the same.

The player's score is calculated as # of simultaneous button presses $\times$ score modifier. "Simultaneous" pressing for this game effectively means that the first pressed key is not released before the second key is pressed. In the game's implementation each individual button press throws an event, which is handled by the OnGUI function such that it can capture multiple events happening as soon as they happen (OnGUI calls as soon as event triggers, and doesn't wait for a frame update). When a key in the correct set is pressed (KeyDown event), the game adds it to the set of keys currently being pressed (CurrentKeys) and flips the key's state to say that it is currently being pressed (isPressed - one variable

Parameter	Default Value	Explanation
Button 1	Right Arrow	The specific button pressed may affect difficulty due to its size and postion.
Button 2	Left Arrow	
Time limit	10 seconds	The time given to the player limits how many times they can press the button. 10 seconds is the default as it reflects the way average tapping speed was measured (60 presses/10 seconds), and is similar in length to many classic button mashing Mario Party minigames.
Goal	60	The expected number of presses is a measure of difficulty for MIBM as this is the "success target". There is no normative data for the number of button sequences pressed during this time. From informal testing with 5 able-bodied people between the ages of 25 and 35 we found the simultaneous presses to be similar in rate to single button pressing.
Score modifier	1	This is a main measure of difficulty as it affects how a button press is counted towards the goal. It provides a way to artificially induce fatigue or skew the difficulty. 1 is the default because it is the intuitive way players will understand their score increasing (1 press = 1 press).

Table D.4: Parameters of MIBM custom game — Fly Away — with the listed default settings and an explanation of the parameter's purpose.

exists per key). On release (KeyUp event), the game removes the unpressed key from the set of currently pressed keys (CurrentKeys) and reverts its state back to unpressed. For the last step of each event (KeyDown or KeyUp) the game checks to see whether the target keys are in the current key set and that no other keys are also being pressed (this is to stop players from just slamming their hand on the whole keyboard). If this is the case the game increases the player's score. Since the score increase is happening through the event system, holding the keys down will not increase the player's score because it will not trigger the events to increase score. Having the scoring happen through the event system also means that players who are effectively alternating (i.e. releasing the first button before the second is pressed) would have a lower score, while rewarding players who are abiding by the coordinated action even if they are moving slower. By designing and implementing the game this way, performance hinges on synchronous pressing (which is in line with the challenge type) in a way that may penalize fast but unsynchronized players (who are effectively alternating key presses) over slow but synchronized players. Given our work sees alternating and synchronous pressing as inherently different, this seems appropriate.

The final score determines how high the witch flies, details can be seen in Table D.5. The values and animations are arbitrary. Since we are working with the final scores for our analysis, this animation and ranking is exclusively to provide feedback to players about their performance.

Rank	Player final score	Flight Animation	Meaning
F	$\leq 0.5*$Goal	Witch falls on her butt	Really bad
D	$0.5*$Goal $\geq$ Score $\leq 0.75*$Goal	Witch hovers just above the ground in the same starting screen	Bad
C	$0.75*$Goal $\geq$ Score $\leq$ Goal	Witch moves into the next screen still in the clouds	Low Average
B	Goal $\geq$ Score $\leq 1.25*$Goal	Witch moves past the clouds into a light blue sky	High Average
A	$1.25*$Goal $\geq$ Score $\leq 1.5*$Goal	Witch moves into the stratosphere with the moon behind her	Good
S	$\geq 1.5*$Goal	Witch escapes orbit and is in the stars	Really good

Table D.5: List of animation effects based on player final score relative to the goal number.

D.0.3 Alternating Input Button Mashing: Potion Master

Figure D.46: Alternating input button mashing challenge designed to be customizable.

Potion Master (Fig. D.46) has our witch brewing a magic potion. The player does this by rapidly pressing two target buttons in a specific repeating sequence (Button 1 then Button 2) within the allotted time. The parameters for the game, along with their default values are found in Table D.6.

The target goal is set to 50, based on our two rounds of internal testing using analog and digital measures. In the first round we asked a random sample of 15 people to perform the sequence as many times as possible in 10 seconds on handheld tally counters (i.e. alternating between right and left hands), resulting in an average of 45 sequences completed in 10 seconds. The second round had our pilot testers play the game and found an average of 50 sequences per 10 seconds.

The player's score is determined by # correct sequences $\times$ score modifier. The sequence order is fixed, as players must always start with Button 1. On correct input, the witch will

Parameter	Default Value	Explanation
Button 1	Right Arrow	The specific button pressed may affect difficulty due to its size and postion.
Button 2	Left Arrow	
Time limit	10 seconds	The time given to the player limits how many times they can press the button. 10 seconds is the default as it reflects the way average tapping speed was measured (60 presses/10 seconds), and is similar in length to many classic button mashing Mario Party minigames.
Goal	45	The expected number of presses is the main measure of difficulty for AIBM as this is the "success target". There is no normative data for the number of button sequences pressed during this time. From informal testing with 5 able-bodied people between the ages of 25 and 35 we found an average of 45 complete sequences.
Score modifier	1	This affects how a button press is counted towards the goal. It provides a way to artificially induce fatigue or skew the difficulty. 1 is the default because it is the intuitive way players will understand their score increasing (1 press = 1 press).

Table D.6: Parameters of AIBM custom game — Potion Master — with the listed default settings and an explanation of the parameter's purpose.

complete a half-stir animation (moving forward on Button 1, and backwards on Button 2). To enforce the patterned input, the game remembers the current position in the sequence and what the next input should be. Only input that matches the next input in the sequence will result in animation and work towards increasing the player's score. This means players who cannot maintain the sequence are not penalized by having to restart the sequence and can use the visual feedback (stirring animation) to know what the "next move" is, but if they are just hitting random keys they will not have a high score. The final score will determine whether how well the potion is brewed.

The results will be displayed through different animations as per Table D.7. These values are fairly arbitrary as they are meant to provide feedback to the player and encourage their play. Since the game outputs raw scores for our analysis we do not need to be specific about this subdivision.

Rank	**Player final score**	**Magic Effect State**	**Effect Meaning**
F	$\leq 0.5*$Goal	Souls Escaping Cauldron	Really bad
D	$0.5*$Goal $\geq$ Score $\leq 0.75*$Goal	Poison Cloud from Cauldron	Bad
C	$0.75*$Goal $\geq$ Score $\leq$ Goal	Purple Smoke Puff	Low Average
B	Goal $\geq$ Score $\leq 1.25*$Goal	Glow	High Average
A	$1.25*$Goal $\geq$ Score $\leq 1.5*$Goal	Electricity	Good
S	$\geq 1.5*$Goal	Magic wall of flames	Really good

Table D.7: List of animation effects based on player final score relative to the goal number.

Appendix E

Competency and Jutsu Validation Experiment Surveys

E.1 Recruiting Participants

Participants were recruited through:

- targeted e-mails to previous participants;
- targeted e-mails to McMaster departments;
- flyers posted on McMaster University's main campus;
- social media posts; and,
- snowball sampling.

We created visual advertisements following the MREB guidelines. Fig. E.47a is for posting in-person across campus. Fig. E.47b is for posting on social media.

Contact Information

Please enter your contact information and Sasha will reach out to eligible participants with more details.

Potential participants are screened via a single screening question.

2. Please enter your preferred name: *

We had unprecedented response to our study. We had over 100 respondents in the first couple hours with just a single e-mail to the Computing and Software undergraduate and graduate mailing lists. The posters had not yet been put up around campus before we had to close the screening survey because we had reached capacity.

E.2 Pre-Study Gaming Habits Survey

3. Please enter your e-mail address *

abilities that relate to performance (i.e. high scores) in different types of button mashing gameplay. Information gathered during this study will be written up as a doctoral thesis, and

This screening survey should take less than 5 minutes to complete.

Eligible participants will then be contacted via the provided e-mail to schedule an in-lab session. **The in-lab session will take a total of 1 hour**. It will involve completing a demographic survey, playing some minigames, and completing an exit survey. **Participants will be compensated $20 for their in-lab time. Participants who withdraw mid-way through the session will be compensated at a rate of $10/30 minutes to the maximum of $20.**

To learn more about this study, particularly in terms of any risks or harms associated with the study, how confidentiality and anonymity will be handled, withdrawal procedures, incentives that are promised, and how to obtain information about the study's results please **read the Letter of Information at https://mcmasteru365-my.sharepoint.com/:b:/r/personal/sorainsm_mcmaster_ca/Documents/ButtonMashingExperiment/Documentation/LOI%20-%20Competency%20Profiles%20for%20Video%20Games.pdf?csf=1&web=1&e=b3Mc5W**

This study has been reviewed and cleared by the McMaster Research Ethics Board (MREB# 6488). If you have any concerns or questions about your rights as a participant or about the way the study is being conducted, please contact:

McMaster Research Ethics Board Secretariat

Competency profiles for video games

This study is being conducted by Sasha Soraine (Principal Investigator) and Dr. Jacques Carette (Faculty Supervisor) of the Department of Computing and Software at McMaster University. They can be reached via e-mail at sorainsm@mcmaster.ca, and carette@mcmaster.ca respectively.

The purpose of the study is to identify and measure the specific human cognitive and motor abilities that relate to performance (i.e. high scores) in different types of button mashing gameplay. Information gathered during this study will be written up as a doctoral thesis, and used in conference and/or journal papers. **There are no specific inclusion criteria for this study.**

This survey should take approximately 5-10 minutes to complete. Participants will then be contacted to schedule an in-lab sessions. The in-lab sessions will take a total of 1 hour. After completing the session, participants will be compensated $20 for their time.

To learn more about this study, particularly in terms of any risks or harms associated with the study, how confidentiality and anonymity will be handled, withdrawal procedures, incentives that are promised, and how to obtain information about the study's results please read the Letter of Information.

This study has been reviewed and cleared by the McMaster Research Ethics Board (MREB# 6488).

If you have any concerns or questions about your rights as a participant or about the way the study is being conducted, please contact:
McMaster Research Ethics Board Secretariat
Telephone: 1-(905) 525-9140 ext. 23142
E-mail: mreb@mcmaster.ca

* Indicates required question

1. Having read the previous preamble OR the linked Letter of Information, I understand that by clicking the "Yes" option below, I agree to take part in this study. *

   Mark only one oval.

   - ◯ YES, I agree to participate in this study. *Skip to question 2*
   - ◯ NO, I do not agree to participate in this study
     Skip to section 9 (Withdrawing from Study)

Consent Statements

This section outlines specific consent statements that must be answered before participating in the survey. Agreeing or disagreeing to these specific consent statements will not be used to disqualify participants from the study.

2. I agree to allow my study data to be stored and used for future research as described in the Letter of Information *

Mark only one oval.

- ◯ Yes
- ◯ No

3. If I choose to quit the study, I agree to have my responses up to the point of quitting the study retained for use in the research. *

Mark only one oval.

- ◯ Yes
- ◯ No

4. I agree to allow my anonymized study data be uploaded to an open science data sharing platform as part of a publication process *

Mark only one oval.

- ◯ Yes
- ◯ No

5. Please select any that apply to you: *

Check all that apply.

- [ ] I would like a copy of the study results sent to the contact e-mail I provided.
- [ ] I would like to be contacted to participate in further studies for this research program.

Contact Information

The following information will be used to contact you in order to set up your study session and optionally to send you copies of the study results. This information will not be stored as part of the study data.

For contact purposes we require an e-mail from participants. Participants who prefer to coordinate over phone (call or text) may provide their contact number, but e-mail is the only required contact method.

Please note that your name and contact information will remain completely confidential and **will not be linked** with any of your study responses.

6. Please enter your full name and preferred pronouns. *

_____

7. Please enter your e-mail address. *

_____

8. What is your preferred methods of contact? Select all that apply.

Check all that apply.

- [ ] E-mail
- [ ] Phone (Call)
- [ ] Phone (Text)

9. Please enter your phone number (for calling).

10. If it is different from the calling number, please enter your phone number for texting.

Demographic Questions

These are questions regarding your personal demographics. We need to collect this data to contextualize the performance data collected during the experiment. If you are uncomfortable providing this information you have the right not to answer; for any required questions (with the red star to the right) you can select or type in "prefer not to answer" and still be eligible for the study.

11. What is your age in years? *

12. What is your **self-identified** gender? *

13. What is your highest level of completed education? *

Mark only one oval.

- ◯ No formal education
- ◯ Some elementary school (Kindergarten to Grade 8)
- ◯ Completed Elementary School (Completed Grade 8)
- ◯ Some high school or equivalent (Grade 9 - 12)
- ◯ Completed High School Diploma OR GED
- ◯ Some college or university
- ◯ Completed College Diploma or University Degree
- ◯ Some post-graduate (Masters or PhD)
- ◯ Completed Post-graduate (Masters or PhD)

14. Do you have any conditions, like carpal tunnel syndrome or repetitive stress injuries, that may impact your ability to use a keyboard for a period of up to 20 minutes? *

Mark only one oval.

- ◯ Yes
- ◯ No

15. Do you **self-identify** as a person with a disability that was present at birth, caused by an accident, or developed over time; *

• that encompasses any degree of physical disability, mental or developmental disability, sensory disability, learning disability, mental health / psychiatric disability, addiction, and life-threatening allergies;

• that may affect full participation in society (school / work);

• that may have been accommodated in workplace / school because of functional limitation as a result of the disability; or,

• who, as a result of self-perception, perception of others, environmental barriers, inaccessible attitudes, or a any combination of these factors, may experience unequal opportunity to access services by reason of the disability?

Mark only one oval.

- ◯ Prefer not to answer *Skip to question 18*
- ◯ No *Skip to question 18*
- ◯ Yes *Skip to question 16*

16. If you identify as having a disability, does your disability affect any of the following? Please select all that apply.

Check all that apply.

- ☐ Vision
- ☐ Hearing
- ☐ Fine motor control (e.g. fingers, toes)
- ☐ Gross motor control (e.g. legs, arms)
- ☐ Attention
- ☐ Short term memory
- ☐ Long term memory
- ☐ Executive functioning

17. Please list any tools or assistive devices you use when playing games, if any.

_____

Video Game Playing Habits

These are questions about your video game playing habits and preferences, if any. They are used to give us insight into how you approach the study. **You do not need to be a gamer to participate in the study.**

18. Do you consider yourself a gamer? *

Mark only one oval.

- ◯ Yes
- ◯ No
- ◯ Other: _____

19. Approximately how long have you been playing video games? *

Mark only one oval.

- ◯ I do not play video games *Skip to question 23*
- ◯ Less than a year
- ◯ Between 1 and 4 years
- ◯ Between 5 to 9 years
- ◯ 10 or more years

20. On average, how many times a week do you play video games on any platform (e.g. phone, tablet, console, computer)

Mark only one oval.

- ◯ Once a week
- ◯ 2 to 4 times a week
- ◯ 5 or more times a week
- ◯ I do not play video games regularly

21. On average, how much time do you spend on a single play session (from turning the game on to turning it off)?

Mark only one oval.

- ◯ 5 to 10 minutes
- ◯ 15 to 30 minutes
- ◯ 30 minutes to 1 hour
- ◯ 1 - 3 hours
- ◯ More than 3 hours

22. What are the games you play (or used to play) frequently? Feel free to give specific titles or just genres you enjoy.

_____

_____

_____

_____

_____

23. What devices do you commonly use? Please select all that apply. *

Check all that apply.

- [ ] Computer
- [ ] Smartphone or tablet
- [ ] Handheld gaming console (e.g. Playstation Vita, Nintendo 3DS)
- [ ] Gaming console (e.g. Xbox One, Playstation 5, Nintendo Switch)
- [ ] Other: _____

24. On a scale of 1 to 5, how comfortable are you with using a keyboard and mouse? *

Mark only one oval.

	1	2	3	4	5	
Very	◯	◯	◯	◯	◯	Very comfortable

25. From the list below, please select any video game controllers that you have used in the past. *

Check all that apply.

- [ ] Keyboard
- [ ] Mouse
- [ ] Joystick and buttons
- [ ] Gamepad (e.g. Xbox controller, Playstation controller)
- [ ] Motion controller (e.g. Wii Remote, Switch Joy-Cons, Playstation Move)
- [ ] Full motion controller (e.g. Kinect)

Survey Complete!

Thank you for completing the pre-study questions. Your responses have been submitted. Our researchers will check your age and gameplaying habits to see if you are eligible for the study before contacting you to set up a session.

If you have any follow up questions please feel free to contact Sasha Soraine by phone (289-434-4053)or via e-mail (sorainsm@mcmaster.ca).

Withdrawing from Study

Thank you for your time. You have decided to quit this study. None of your responses have been collected or stored.

E.3 Post-Jutsu Experiement Survey

Post Button Mashing Survey

This survey will ask you questions about your experience with the button mashing games you have just played.

* Required

1. The scales in this question run from 1 (Strongly disagree with the statement) to 7 (Strong agree with the statement); a 4 on this scale is neither agree nor disagree. *

	Strongly Disagree	Disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Agree	Strongly agree
Playing the game was meaningful to me.	◯	◯	◯	◯	◯	◯	◯
The game felt relevant to me.	◯	◯	◯	◯	◯	◯	◯
Playing this game was valuable to me.	◯	◯	◯	◯	◯	◯	◯
I felt I was good at playing this game.	◯	◯	◯	◯	◯	◯	◯
I felt capable while playing the game.	◯	◯	◯	◯	◯	◯	◯
I felt a sense of mastery playing this game.	◯	◯	◯	◯	◯	◯	◯
I was no longer aware of my surroundings while I was playing.	◯	◯	◯	◯	◯	◯	◯
I was immersed in the game.	◯	◯	◯	◯	◯	◯	◯
I was fully focused on the game.	◯	◯	◯	◯	◯	◯	◯
The game							

informed me of my progress in the game.	○	○	○	○	○	○	○
I could easily assess how I was performing in the game.	○	○	○	○	○	○	○
The game gave clear feedback on my progress towards the goals.	○	○	○	○	○	○	○
I enjoyed the way the game was styled.	○	○	○	○	○	○	○
I liked the look and feel of the game.	○	○	○	○	○	○	○
I appreciated the aesthetics of the game.	○	○	○	○	○	○	○
The game was not too easy and not too hard to play.	○	○	○	○	○	○	○
The game was challenging but not too challenging.	○	○	○	○	○	○	○
The challenges in the game were at the right level of difficulty for me.	○	○	○	○	○	○	○
It was easy to know how to perform actions in the game.	○	○	○	○	○	○	○

The actions to control the game were clear to me.	◯	◯	◯	◯	◯	◯	◯

2. The scales in this question run from 1 (Strongly disagree with the statement) to 7 (Strong agree with the statement); a 4 on this scale is neither agree nor disagree. *

	Strongly Disagree	Disagree	Somewhat disagree	Neither agree nor disagree	Somewhat Agree	Agree	Strongly agree
I thought the game was easy to control.	◯	◯	◯	◯	◯	◯	◯
I thought the game was easy to control.	◯	◯	◯	◯	◯	◯	◯
I grasped the overall goal of the game.	◯	◯	◯	◯	◯	◯	◯
The goals of the game were clear to me.	◯	◯	◯	◯	◯	◯	◯
I understood the objectives of the game.	◯	◯	◯	◯	◯	◯	◯
I liked playing the game.	◯	◯	◯	◯	◯	◯	◯
The game was entertaining.	◯	◯	◯	◯	◯	◯	◯
I had a good time playing this game	◯	◯	◯	◯	◯	◯	◯

3. What did you feel was the hardest part of the fire starting game? *

4. What did you feel was the hardest part of the potion making game? *

5. What did you feel was the hardest part of the flying game? *

6. Are there any comments you would like to leave the researchers about your experience with the games? *

(a) (b)

Figure E.47: MREB approved recruitment visual materials; E.47a is approved for physical posting, E.47b is approved for posting on social media.

Appendix F

Background Information About Methods for Study

Multiple linear regression (MLR) finds a linear combination of multiple predictor variables that provides optimal explanatory power for a single dependent variable. In terms of our work, MLR models attempt to quantify how the given abilities affect in-game performance. MLR requires that we, as domain experts, provide it with predictors we are confident relate to the dependent variable. Since we are unsure which predictors belong in the models, we take a more exploratory approach and use subset analysis procedures.

F.0.1 Subset Analysis: Stepwise Regression

Subset analysis procedures are strategies for MLR that explore the data in an attempt to find a subset of predictors that optimizes the model in someway [168]. There are two general approaches to finding a subset model: *all-subsets regression* and *stepwise regression* [168]. *All-subsets regression* computes fitted models for all possible subsets of the predictors. These models are compared based on a calculated criterion[1] to create a short list of potential regressions that would be further examined for suitability. *Stepwise regression* starts with an initial model (either without any predictors, or with all predictors included) that is refined in steps that add or remove a variable based on inclusion and exclusion criteria (most frequently F-statistic thresholds on a regression diagnostic or fit metric). Stepwise regression is notably a more widely used method, likely due to its inclusion in common statistical software packages and lower computational cost [**<empty citation>**].

While frequently used, stepwise regression is significantly criticised [185]. Concerns range from fundamental statistical issues with the procedure (e.g. bias of least squares regression), lack of academic rigor (e.g. not assessing the underlying assumptions), and misuse (e.g. researchers using it in place of theory-driven models) and misreporting (e.g. reporting resulting models as "best" or "correct") in academic work. Many critics note that stepwise regression should only be considered acceptable for small, exploratory studies, where little-to-no theory exists or where researchers are looking for insight into a hypothesized theory [185]. This constraint allows for starting inquiries into areas and the small size reduces some of the statistical bias concerns.

[1] Common criteria are R^2, Adjusted ^{2}, root mean square of residuals, or Mallow's C_p [168]

Our work is exploratory and lacks a well-established theory-driven model, making it appropriate for stepwise regression. Given we do not have a solid reason to assume a starting model, we start our stepwise regression with an empty model (i.e. just a constant). We then allow the process to add or remove terms based on whether they created a significant improvement in the model's sum of squared error (SSE)[2]. The SSE is a measure of the model's estimation power based on the difference between the real values and model's predicted values. A smaller SSE means the regression fits the data better, and gets us closer to a more precise model. A model is judged as "better" if the addition/removal of a term results in a significant result for an F-Test comparing the two models' SSE. Given that this is a first exploration of competency profiles, we set the addition/removal threshold to $p < 0.05$. The resulting reduced model is then checked against the common assumption and diagnostics for multiple linear regression to assess whether they are "reasonable" (details in Sec. F.0.2).

F.0.2 Assumptions of Regression Modeling

For a multiple linear regression model to be reasonable, it needs to meet the following assumptions:

1. each independent variable is correlated to the dependent variable;
2. the independent variables are not collinear with each other;
3. the errors follow a normal distribution;
4. the errors are independent; and,
5. the errors are homoscedastic.

We can think of these assumptions having to do with the independent variables (Assumptions 1 and 2) or having to do with the randomized errors of the regression (Assumptions 3, 4, and 5). The following subsections cover these assumptions in more detail.

At this point we want to note that the software we will be using to measure the player's abilities (independent variables) is not a perfect measurement tool (see Ch. 11). While for the purposes of this study we will consider the measurements as "real", we need to consider how the errors in ability measurement may influence the assumptions and regression. Generally this could introduce noise into the data such that strict interpretations of assumptions would not be met. We keep this in mind when evaluating the assumptions, and explain below how each will address this potential noise.

Assumptions on Independent Variables

Assumptions 1 and 2 can be checked prior to running the stepwise regression.

[2]We select SSE over R^2 because R^2 will generally increase with more predictors in the model, thus biasing the resulting model.

Correlation. The purpose of checking for correlation is to narrow down which independent variables should reasonably occur in the final regression model. As the point of this is to setup for regression, we do not put a threshold for correlation coefficients (Pearson's r) at this time. However, we do care about significance of the correlation and choose to set $\alpha = 0.05$. Given the exploratory nature of this work, and its gaming context, a stricter α than this traditional baseline did not seem necessary. Any independent variables that are not significantly correlated with the dependent variable should not be in the regression model.

Collinearity. We check for collinearity to see whether our model's coefficients and included variables are being affected by relationships between our variables (as this could make it harder to interpret the model). We check for collinearity of our variables with a pairwise correlation matrix and variance inflation factors (VIF). We consider correlations and collinearity within acceptable ranges if the correlation coefficient between any two independent variables is $r < 0.5$. This value is somewhat arbitrary, but we contextualize it with the VIF values for the variables. Variables are considered not correlated when their VIF is equal to 1, moderately correlated for values between 1 and 5, and highly problematically correlated for values above 5. Since human cognitive and motor abilities tend to be redundant (e.g. finger pressing measured in all tasks because it is the motor response) and interconnected (e.g. selective attention and inhibition), we expect some amounts of correlation. Since we know we will not have clean orthogonal ability measures, we have looser constraints around acceptable correlation to account for this. We consider VIF values less than 3 to be acceptable since we expect some amount of correlation in the variables (particularly the cognitive measures) — this follows a commonly cited "rule of thumb" for judging problematic multicollinearity. While we will report the VIF values for all variables together, we will also report VIFs when resulting regression models include multiple predictors to show how the collinearity looks for that subset.

Assumptions on randomized errors in model

Assumptions 3, 4, and 5 can only be checked after the stepwise regression is run. We evaluate all of these assumptions through analysing the residuals for the model. Given how the potential measurement errors in the independent variables would propagate in the regression, we evaluate the assumptions for residuals visually to get a sense for the severity of assumption deviations. We outline below the deviations we may expect to give readers a sense of what we believe would be acceptable deviations.

Residuals are normally distributed. We check the normality of the residuals through looking at their histogram, probability (P-P) plot, and quantile-quantile (Q-Q) plot.

We expect the residuals to follow a normal distribution considering we are looking at the skill-performance relationship of a random sample of the general population. However, since the skill-performance relationship is also influenced by practice and age, we could see some skew based on the demographics of participants. Participants from a university-aged cohort are likely at the peak of age-related effects, and gamers are likely at the peak of practice effects compared to non-gamers. This could create tailing in our data and residuals which

would show up in these normality plots. So long as the normality plots do not deviate in unexpected ways, we can say the normality assumption is met.

Residuals are Independent and Homoscedastic. We check for independence and homoscedasticity in the residuals visually through a residuals versus fitted values plot. For these plots, we are looking for the data points to be randomly distributed with no specific patterns (independent) and for the data points to be relatively symmetrical around the 0-line (homoscedastic). We expect that the same skew from particularly "skilled" participants will cause some extreme outliers, but so long as no obvious pattern exists we believe these are acceptable.

Appendix G

Over and Under-loading Study Details

G.1 Checking Regression Assumptions

G.1.1 Easy Condition

We start by taking a look at correlations between independent variables and the dependent variables, summarized in Table G.8.

	Correlations				
	FP	**SelAtt**	**SIBM**	**MIBM**	**AIBM**
FP	–	-0.1283	0.7536****	0.5841***	0.5392**
SelAtt	-0.1283	–	-0.1714	-0.0263	0.0837

Significance: not significant ( ), $p < 0.05$(*), $p < 0.01$(**), $p < 0.005$(***), $p < 0.001$(****)

Table G.8: Summary of Correlations for Easy Condition. FP: Finger Pressing, SelAtt: Selective Attention, SIBM: Single Input Button Mashing, MIBM: Multiple Input Button Mashing, AIBM: Alternating Input Button Mashing.

SIBM

The SIBM regression model (summarized Table 17.2 and Fig. 17.8), is significant. The single predictor variable used (Finger Pressing), is significantly correlated to the SIBM Scores (Tbl. G.8). Looking at its Residuals versus Fitted values (Fig. G.48) shows a fairly random distribution of points. There seems to be some clustering above the 0-line, but with so few points it does not seem like these indicate a larger latent pattern or trend. This could be happening because the data itself is skewed, and therefore the residuals may similarly be skewed from normal.

Looking at the P-P Plot (Fig. G.49a) we see some bulging around the centre that almost looks s-curved. The Q-Q Plot (Fig. G.49b) shows similar s-curvature. This seems to indicate that data is peaky in the middle, and had long tails which pull the values. Overall the values do not actually deviate drastically from the reference line is a way that obviously indicates a

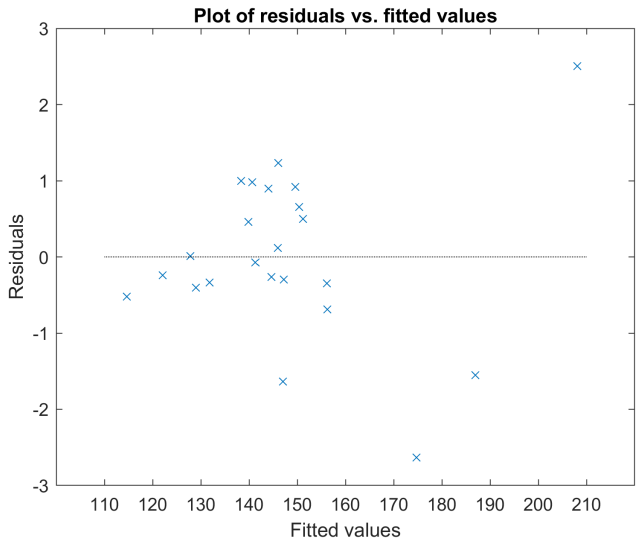
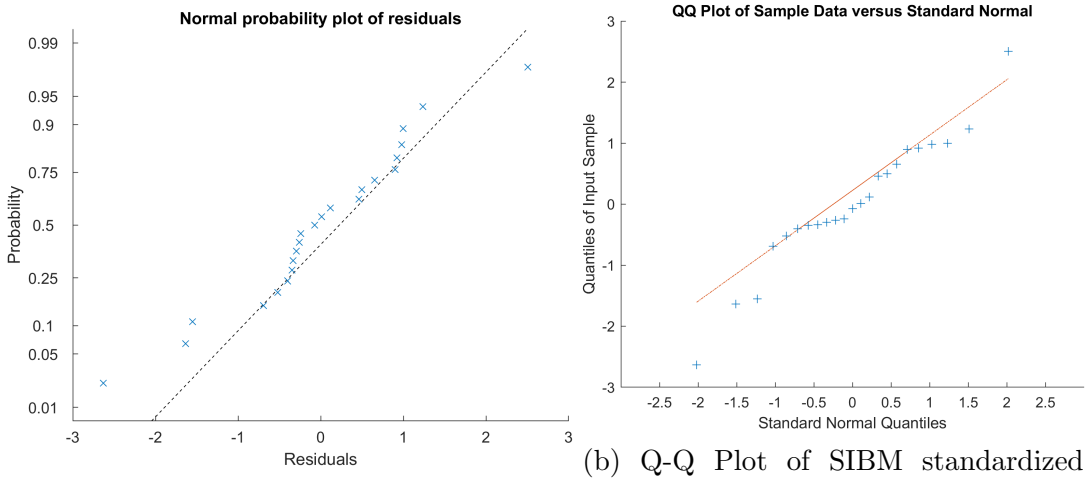


Figure G.48: Standardized Residuals vs. Fitted Values Plot for SIBM (Easy).

particular non-linear relationship. It could be that the data is non-linear, or more likely we do not have a sufficient number of points to approach the normal distribution.



(a) P-P Plot of SIBM standardized residuals with Normal Distribution reference line.

(b) Q-Q Plot of SIBM standardized residuals with Normal Distribution reference line.

Figure G.49: Normal probability plots for SIBM Standardized Residuals.

Checking the histogram of residuals (Fig. G.50) we see heavy tails and peaking. However, the plot does not seem to significantly deviate from normal, nor does it immediately call to mind another distribution that would better fit the data.

Overall, given the small sample size and the exploratory nature of this study, the deviations from normality assumptions seem reasonable. We think this indicates the resulting regression model is reasonable.

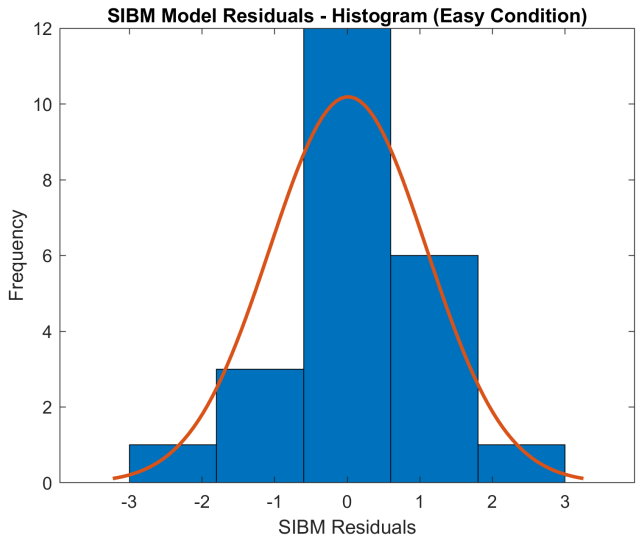


Figure G.50: Histogram of SIBM standardized residuals with a normal distribution reference line.

MIBM

The MIBM regression model (summarized in Table 17.4 and Fig. 17.10) is significant. The single predictor variable (Finger Pressing) is significantly correlated to the score. Looking at the residuals vs fitted values (Fig. G.51) shows points on either side of the 0-line. It is difficult to tell from the few points we have whether the residuals form an obvious pattern. There is some linear growth in points above the 0-line ending in a data cluster. There is also some increasing and decreasing in the bottom half of the graph that may indicate a quadratic-style curve. However, it is hard to tell with the few number of points we have whether this is over-interpreting this plot.

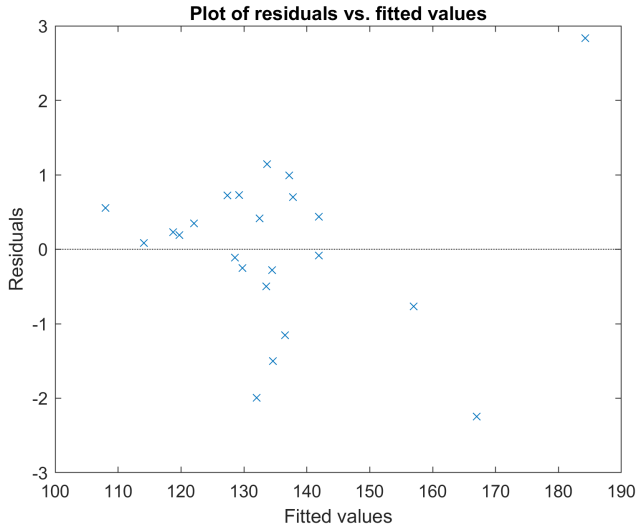
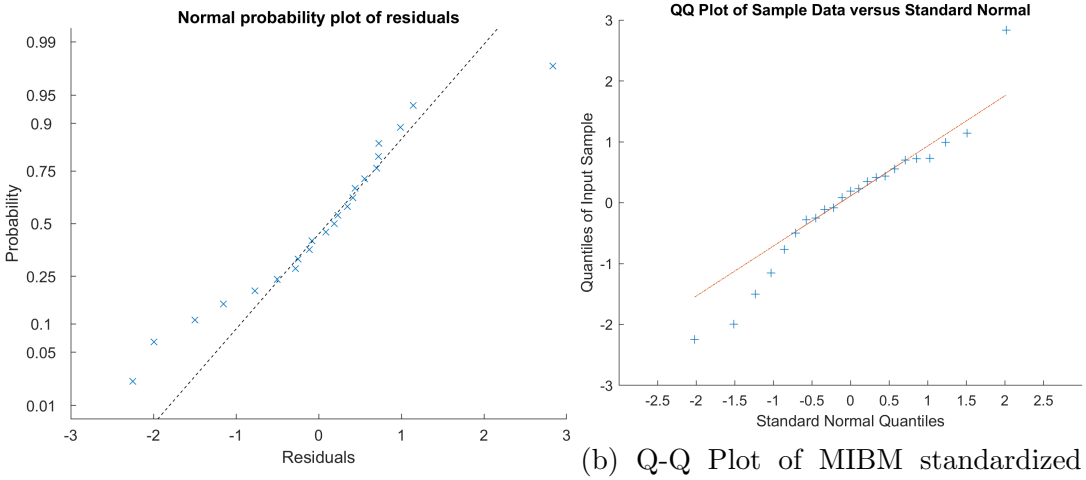


Figure G.51: Standardized Residuals vs. Fitted Values Plot for MIBM (Easy).

Turning to the P-P Plot (Fig. G.52a) and Q-Q Plot (Fig. G.52b) we see that the line is

quite normal in the centre of the data but skews in the negative residuals. There seems to be a wide spread of values at the tail, more so than expected, and the drastic shapes indicate heavy tailing.



(a) P-P Plot of MIBM standardized residuals with Normal Distribution reference line.

(b) Q-Q Plot of MIBM standardized residuals with Normal Distribution reference line.

Figure G.52: Normal probability plots for MIBM Standardized Residuals.

The residuals histogram (Fig. G.53) confirms what we are seeing in the probability plot. Overall this looks like it could come from a normal distribution if more data points were gathered.

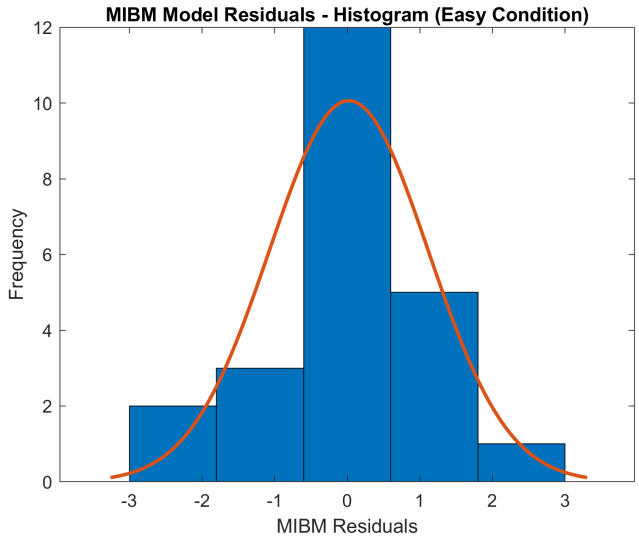


Figure G.53: Histogram of MIBM standardized residuals with a normal distribution reference line.

When considering this information against the regression model plot (Fig. 17.10), this lack of random distribution and heteroscedastic behaviour in the residuals could be the result of the significantly underperforming datapoints in the set. It is unclear why this

underperforming could be happening, and they fit within the prediction interval so they do not deserve to be removed at the moment.

Overall we will tentatively consider this model acceptable, however a larger sample size would likely improve our ability to determine whether this model meets assumptions.

AIBM

The AIBM regression model was significant (summarized in Table 17.6 and Fig. **??**). Looking at the correlations for the predictor variables Finger Pressing is significantly correlated to the scores, but Selective Attention is not. This inherently means it "ought not" to be in the model, however it remains for theory reasons (because it shows up significantly in the baseline).

Taking a look at the residuals vs fitted values plot (Fig. G.54) we see what looks like a quadratic pattern, which would indicate a linear model is not a reasonable way to look at this data. However, given the small sample size for a two-predictor regression we may be over interpreting what could be noise. There is some clustering of data points between 0 and 1, and there are significant outliers around -2 and +3 which could similarly be cause our impression of the data to be skewed.

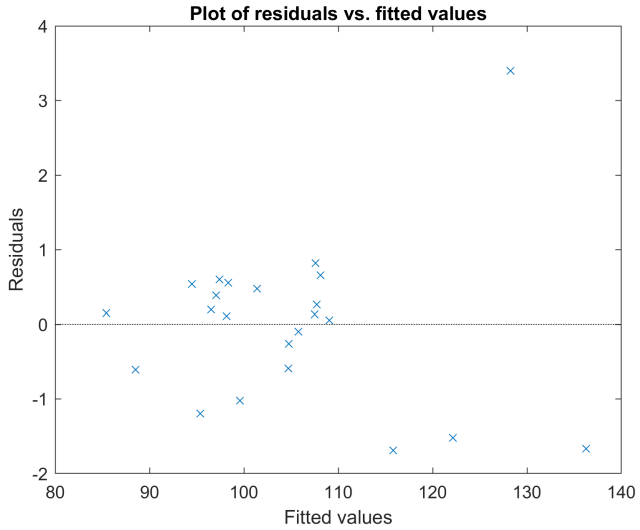
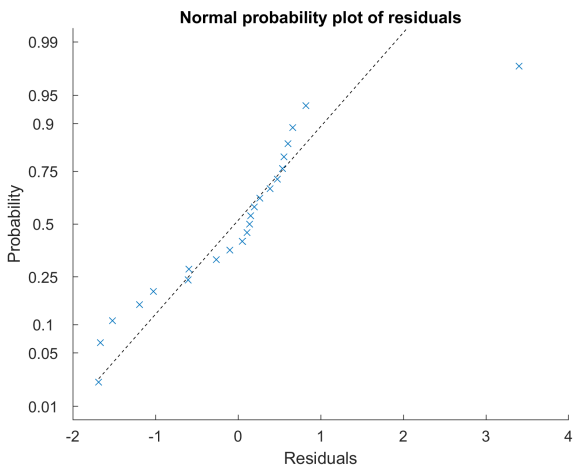


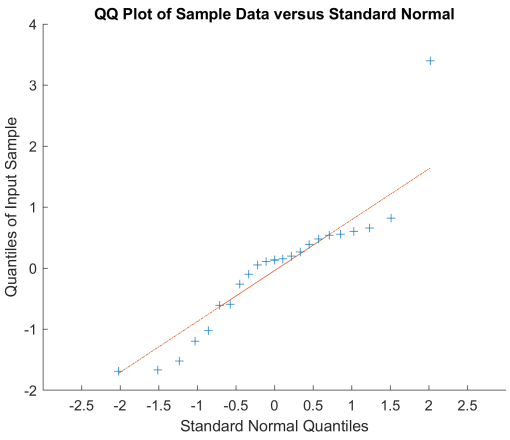
Figure G.54: Standardized Residuals vs. Fitted Values Plot for AIBM (Easy).

Focusing on the P-P Plot (Fig. G.55a) and Q-Q plots (Fig. G.55b) we see significant tailing, bulging around the centre, and an odd spacing for the distribution. The data looks like it is quite shifted from the normal distribution, but again this could be due to an overinterpretation of a small sample size.

The histogram of the residuals (Fig. G.56) shows that the residuals may not be normally distributed. While most of the data shows up under the normal reference line, there is significant bunching around the centre (which matches the bulging in our probability plots). The tailing that we are seeing similarly seems overblown because of the sample size; the bins for the +3 residual has only 1 data point in it, while the bin between -1 and -2 has 5. This histogram leads us to believe that the outliers are having a significant impact on the overall data.



(a) P-P Plot of AIBM standardized residuals with Normal Distribution reference line.



(b) Q-Q Plot of AIBM standardized residuals with Normal Distribution reference line.

Figure G.55: Normal probability plots for AIBM Standardized Residuals.

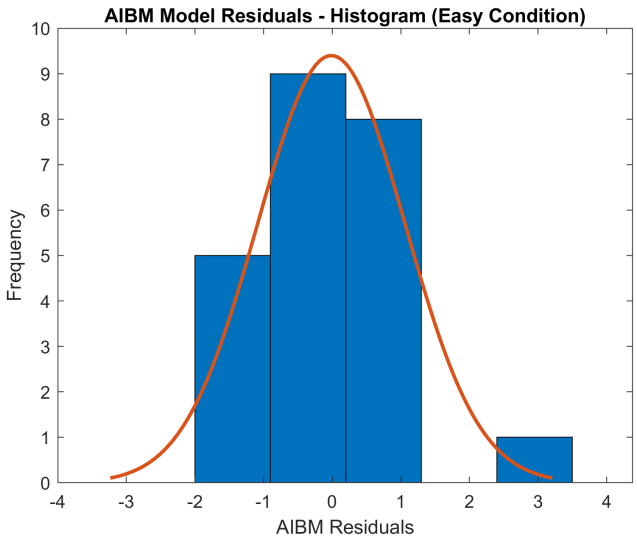


Figure G.56: Histogram of AIBM standardized residuals with a normal distribution reference line.

Overall, we do not consider this model acceptable by strict regression assumptions due to the lack of correlation between Selective Attention and the AIBM scores, the patterning in the residuals vs. fitted plot, and the seemingly non-normal residuals. However, while this model does not meet the regression assumptions, we feel this is more due to the sample size limitations than the underlying theory being incorrect.

G.1.2 Hard Condition

We start by taking a look at correlations between independent variables and the dependent variables, summarized in Table G.9. Notably this is the only condition where Selective

Attention and Finger Pressing significantly correlate with each other. This seems to be an artifact of the participant subset, as these abilities do not correlate in the baseline study where all participants were used.

	Correlations				
	FP	**SelAtt**	**SIBM**	**MIBM**	**AIBM**
FP	–	0.5051**	0.7874****	0.6410****	0.5419**
SelAtt	–	–	0.4974*[1]	0.1522	0.3565[2]

Significance: not significant ( ), $p < 0.05(^{*})$, $p < 0.01(^{**})$, $p < 0.005(^{***})$, $p < 0.001(^{****})$

Table G.9: Summary of Correlations for Easy Condition. FP: Finger Pressing, SelAtt: Selective Attention, SIBM: Single Input Button Mashing, MIBM: Multiple Input Button Mashing, AIBM: Alternating Input Button Mashing.

SIBM

The SIBM regression model was significant (summarized in Table 17.3 and Fig. 17.9). The sole predictor (Finger Pressing) is significantly correlated to the SIBM scores. However, Selective Attention also significantly correlates with the SIBM scores. Looking at the residuals vs. fitted plot (Fig. G.57) we see a very dispersed set of data points. There is no obvious variance change related to the fitted values (i.e. not heteroscedastic), but the dispersion makes it hard to see whether there are any obvious patterns emerging.

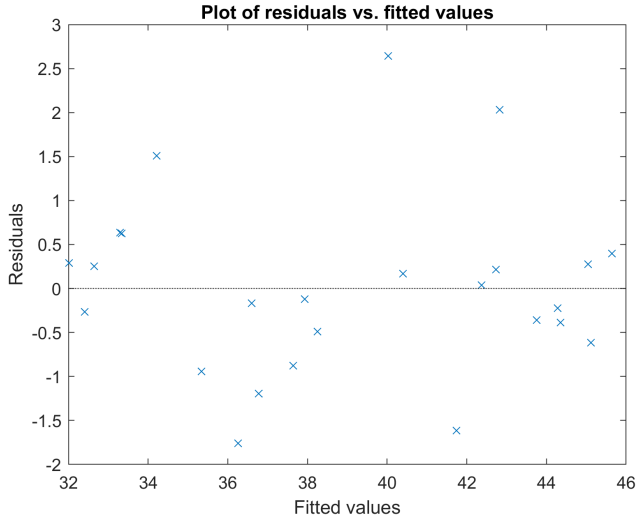
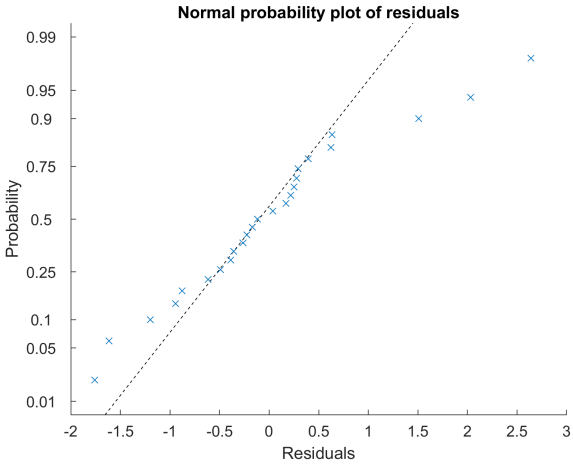


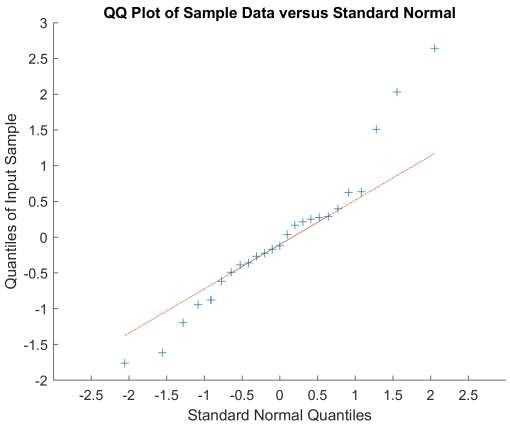
Figure G.57: Standardized Residuals vs. Fitted Values Plot for SIBM (Hard).

Looking at the P-P Plot (Fig. G.58a) we see some bulging in the centre that could indicate peaking around the 0.5 Residuals. Turning to the Q-Q Plot we clearly see heavy tailing as the positive quantiles skew far from the reference.

[1] p=0.0114
[2] p=0.0802



(a) P-P Plot of SIBM standardized residuals with Normal Distribution reference line.



(b) Q-Q Plot of SIBM standardized residuals with Normal Distribution reference line.

Figure G.58: Normal probability plots for SIBM Standardized Residuals.

Looking at the residual's histogram (Fig. G.59) shows strong peaking and tailing on the right.

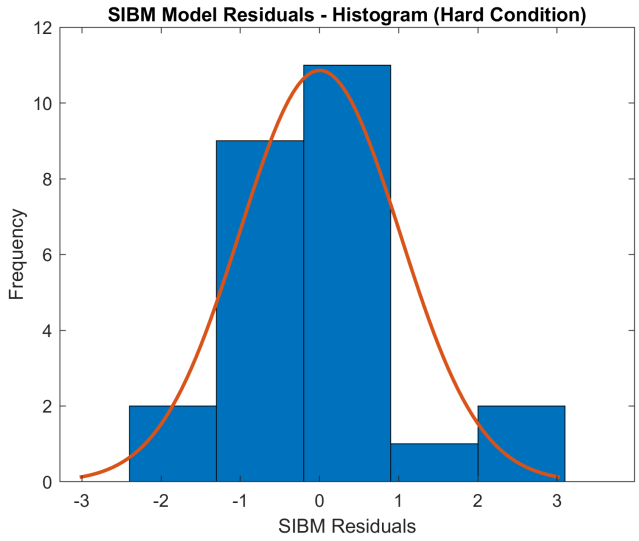


Figure G.59: Histogram of SIBM standardized residuals with a normal distribution reference line.

Looking at these diagnostic plots it seems that we're seeing that the regression is over-predicting player performance. It is possible that a larger sample size may produce more normally distributed residuals and clear elements to support or refute the assumptions. As well, the correlation between SIBM scores and Selective Attention is surprising, and could imply that a multi-variable model would better fit this data. As it stands, we can tentatively say this model is reasonable, though a larger study may show something different.

MIBM

The MIBM model is significant (summarized in Table 17.5 and Fig. 17.11). The only predictor variable (Finger Pressing) is significantly correlated with MIBM scores. Looking at the residuals vs. fitted plot (Fig. G.60) we see some patterning (though this could be an artifact of over interpreting a small amount of data). We see a potential linear pattern for points on the left side of the plot (Fitted values 30 to 36), as well as a potential "W" pattern with points as they bounce around the positive residuals and dip slightly below the 0-line. We also notice that the plot shows significant outliers in the negative side.

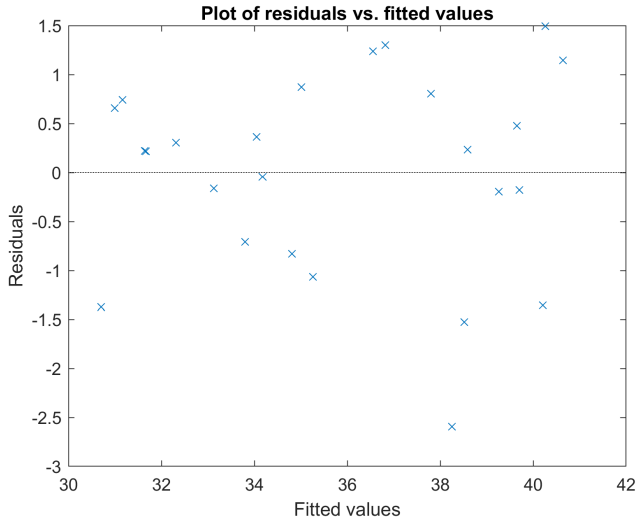
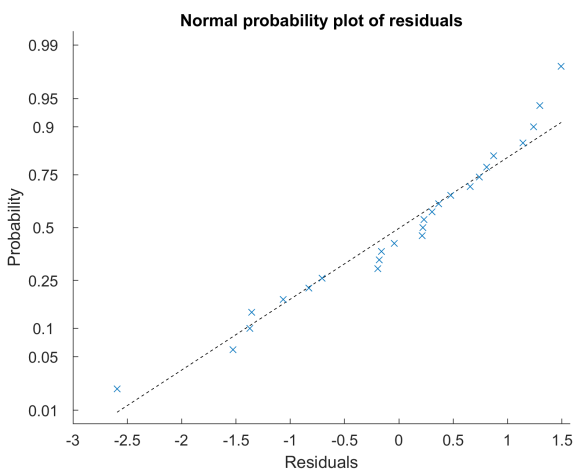


Figure G.60: Standardized Residuals vs. Fitted Values Plot for MIBM (Hard).

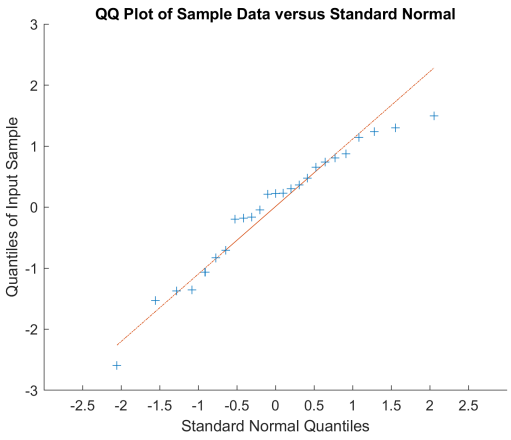
The probability plots (Fig. G.61) show the data points are close to the reference line. However, there are elements of the distributions that make us suspect the residuals may not be normally distributed. Both plots seem to be skewed in the data point locations and space between them. The P-P plot (Fig. G.61a) has unexpected gaps between data clusters that could be indicative of a latent variable. The Q-Q plot (Fig. G.61b) shows the tails are closer together than we would expect, and the tails begin to skew in the positive quantiles indicating a right skewed residual distribution.

Looking at the histogram we see strong left tailing indicating that the regression seems to be underperforming.

Overall it is hard to assess whether this model is reasonable. The plots are fairly good for human data which is noisy and variable. However there seems to be evidence that the residuals are not normally distributed and there may be some latent variable or condition that is being measured here. We see this specifically in the gaps in the probability plots, and the correlation with Selective Attention, which could point to there being a relationship there. To be cautious and in line with what we saw from the Baseline regressions, we do not consider this model reasonable.



(a) P-P Plot of MIBM standardized residuals with Normal Distribution reference line.



(b) Q-Q Plot of MIBM standardized residuals with Normal Distribution reference line.

Figure G.61: Normal probability plots for MIBM Standardized Residuals.

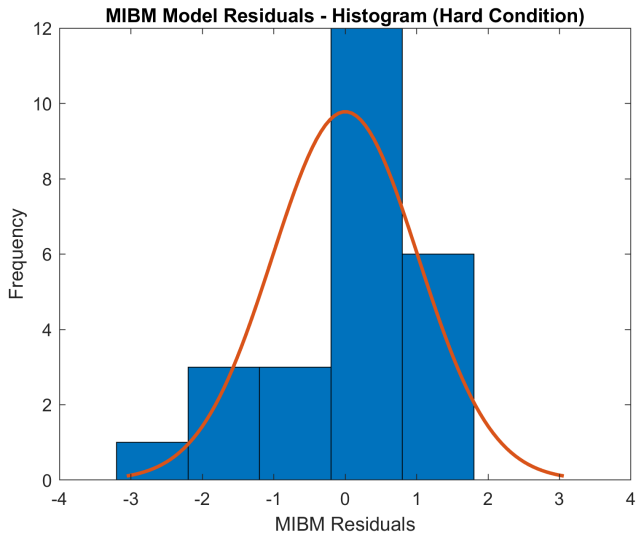


Figure G.62: Histogram of MIBM standardized residuals with a normal distribution reference line.

AIBM

The AIBM regression model was significant (summarized in Table 17.7 and Fig. **??**). Looking at the correlations to AIBM Scores, we see that Finger Pressing (r = 0.5419, p=0.0051) correlates significantly, but Selective attention does not (r=0.3565, p=0.0802). Given how close the measured significance of Selective Attention is to AIBM, it is possible that a larger sample size would result in a significant correlation. While not strictly met, it is reasonable to consider the correlation assumption as practically met given the theory from the baseline models.

Looking at the residuals vs. fitted plot (Fig. G.63) we see the data points are very dispersed, however they seem evenly distributed above and below the 0-line. It seems like

there may be some patterns in the residuals as the points between fitted values 22 and 27 seem to have a downward curve, and the points right of 28 seem to have a slight curve as well. Again, this could be over interpreting information from a graph with too few data points. However, the two distinct sides seems to imply there may be something else at play with this data.

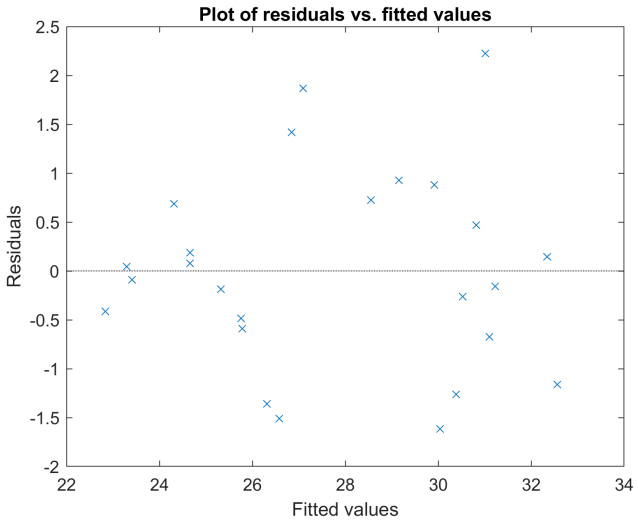
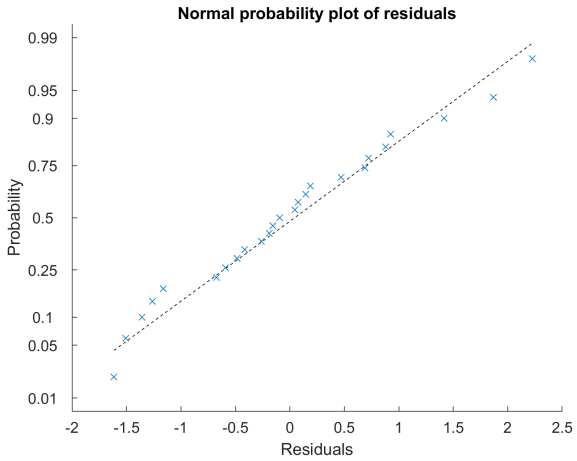
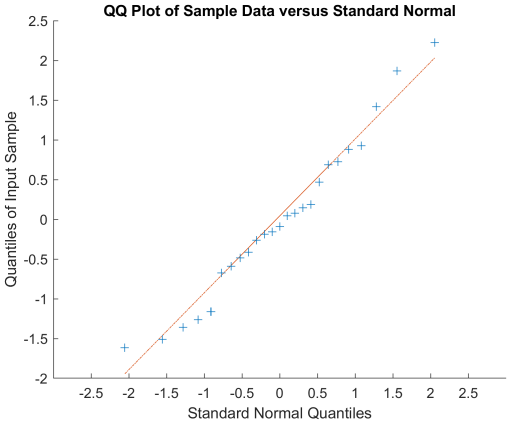


Figure G.63: Standardized Residuals vs. Fitted Values Plot for AIBM (Hard).

Looking at the P-P Plot (Fig. G.64a) we see that the datapoints are close to the reference line, but there are significant gaps between sections, especially around residual -1. The centre of the P-P plot has a couple notable gaps that could be indicating peaking in the data or some sort of missing condition/split in the data. Moving to the Q-Q plot (Fig. G.64b) we see that there is slight tailing and the gap between the centre data and tails is still apparent.



(a) P-P Plot of AIBM standardized residuals with Normal Distribution reference line.



(b) Q-Q Plot of AIBM standardized residuals with Normal Distribution reference line.

Figure G.64: Normal probability plots for AIBM Standardized Residuals.

Looking at the residuals histogram (Fig. G.65) there is significant peaking and a lot of

residuals 2 deviations from the mean. However, overall this looks relatively normal.

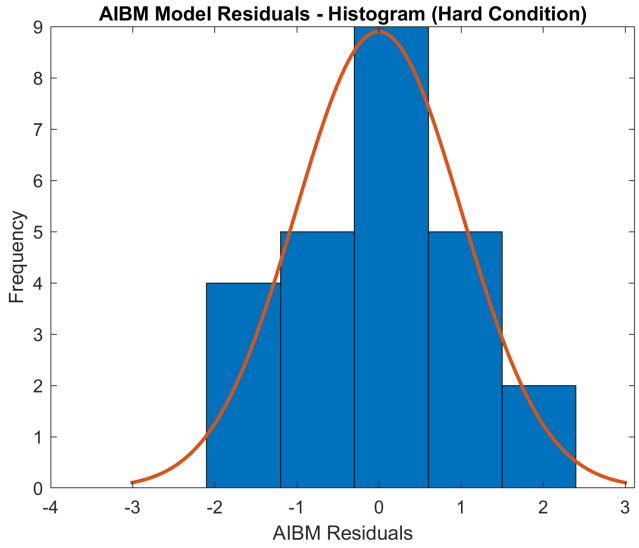


Figure G.65: Histogram of AIBM standardized residuals with a normal distribution reference line.

Overall it seems like this model may have some underlying issues that indicate there is another factor at play not counted in the model. However, the performance of the model seems reasonable, and some of the issues we found are possibly the result of a small sample size. We will cautiously consider this model reasonable, though this may be disproven with a larger sample size.

Appendix H

Mixed Methods Study Details

H.1 Background on Mixed Methods

Mixed methods have been used across many domains since the early 1990s with the motivation of either generating complementary knowledge through multiple data types, or confirming results from one type of investigation with another [437]. The core questions driving mixed methods research designs are *why* are we collecting multiple types of data, and *when* should we be collecting this data. Creswell and Creswell [101] identify three core mixed-methods research designs that address these questions: convergent, exploratory sequential, and explanatory sequential. Convergent designs collect and analyse multiple types of data simultaneously, often for the purpose of contextualizing the gathered quantitative data with the qualitative data for a particular experiment or intervention. Exploratory sequential designs first collect and analyse qualitative data to establish theories and concepts that are then explored using quantitative methods. Explanatory sequential designs first collect and analyse quantitative data, and then follow-up with qualitative methods in order to explain the results further. Since our study is a convergent design, we will focus in on elements related to that design type.

We construct a *questionnaire variant* of convergent design, where the data is collected through a single measure with both open- and closed-ended questions [100]. Creswell and Clark [100] outline the general procedure for convergent designs (including questionnaire variants) as:

1. Collect the datasets;
2. Analyse each data type independently;
3. Integrate the datasets; and,
4. Interpret the merged results.

The independent analysis allows for the analysis methods to be appropriate to each data type, and for preliminary understandings about the data to be formed from these methods. The integration can be done in multiple ways depending on *what* you want to learn from the combination. Since our goal is to understand how the qualitative and quantitative data

compares, we focus on a side-by-side comparison where the results of each data type are first presented independently and then compared in a meta-discussion [101].

Considerations for Convergent Designs. There are two main limitations that could impact our results from this design. Firstly, the questionnaire variant's data collection does not result in rigorous qualitative data since the open-ended questions are just appended to a quantitative survey. While this means the data would not hold up on its own as a qualitative study, it can still produce emergent themes to enhance and discuss the quantitative data set [100]. Since this is suitable for our larger goal, and pragmatically works for our larger experimental session we do not see it being a problem. Secondly, there is tension in finding the right sample size for the study because of the competing rationales from the qualitative (few participants, lots of rich data) and quantitative (lots of participants, select data) perspectives. For our study, since we do not have a rigorous qualitative dataset anyway, we choose to include all of the participants from our quantitative dataset in our qualitative dataset. This will allow for more reasonable comparisons between findings since the groups are covering the same people, even though it means that we are getting a limited amount of qualitative data from each participant.

H.2 Background on Thematic Analysis

Thematic analysis is a method of examining patterns and themes through a dataset for the purpose of finding shared understandings and meanings from the sample population about the concept being researched [59]. Generally, thematic analysis can be broken into a six-stage, non-linear, process [59, 469]:

1. **Familiarization with the data**: becoming immersed in the data through repeated close readings, while keeping notes on elements and patterns that may be useful to analyse.

2. **Initial code generation**: systematically creating and applying labels to the data in an effort to capture its relevant meanings for the research question.

3. **Constructing themes from codes**: grouping related codes in ways that show a cohesive concept or phenomenon.

4. **Reviewing potential themes**: reflexively evaluating themes to see whether they tell a cohesive and coherent story about the data.

5. **Defining and naming themes**: creating illustrative names and descriptions for the final set of themes.

6. **Producing the report**

The specifics of these stages looks different depending on the type of thematic analysis being done. Terry et al. [469] broadly defines two conceptual approaches for thematic analysis: a positivist approach focused on pre-generating codes from theory and testing for the reliability of these codes in the data (the "small q" approach), and a qualitative paradigm approach

focused on code and theme development from exploring and immersing the researcher in the data (the "big Q" approach). The difference between these approaches tends to be their theoretical perspective (constructivist or essentialist), orientation to the data (critical or experiential), and approach to coding and data analysis (deductive or inductive) [59]. While Braun and Clarke [59] note that in practice thematic analysis will often incorporate aspects of each perspective, orientation, and analysis type, it is important that the predominant frameworks are stated to make it clear what can or cannot be said about the data, and what assumptions underlie the interpretations.

We outline our theoretical and conceptual frameworks, and address the underlying assumptions in our analysis as well as our positionality[1] as researchers in Sec. 18.1. We cover our specific implementation of the steps and analysis in Sec. 18.5.1.

H.3 Participant Data

H.3.1 Participant Demographics

Table H.10: Participant Demographics.

	Demographics		Gaming Information			
ID	**Age**	**Gender**	**Gamer**	**History**	**Length (hours)**	**Sessions per week**
P01	30	Man	Yes	10+ years	0.5 to 1	5+
P02	21	Man	Yes	10+ years	1 to 3	5+
P03	28	Man	Yes	10+ years	1 to 3	2 to 4
P04	21	Man	Yes	10+ years	0.25 to 0.5	5+
P05	18	Woman	Yes	1 to 4 years	3+	5+
P06	20	Man	Yes	10+ years	1 to 3	2 to 4
P07	22	Man	No	10+ yeas	1 to 3	Irregular
P08	20	Non-binary	Yes	10+ years	0.5 to 1	5+
P09	20	Man	Yes	10+ years	1 to 3	5+
P10	20	Man	Yes	10+ years	1 to 3	5+
P11	22	Man	Yes	10+ years	1 to 3	2 to 4
P12	20	Man	Yes	10+ years	3+	5+
P13	19	Man	Yes	5 to 9 years	1 to 3	5+
P14	19	Woman	No†	1 to 4 years	Irregular	Irregular

†: indicates a more complicated response than Yes/No

[1]Our relationship to various social processes and hierarchies of power and privilege that impact our approach to the research and insight into the data [383].

Table H.10 (continued)

	Demographics		Gaming Information			
ID	**Age**	**Gender**	**Gamer**	**History**	**Length (hours)**	**Sessions per week**
P15	22	Man	Yes	5 to 9 years	1 to 3	1
P16	22	Man	Yes	5 to 9 years	1 to 3	2 to 4
P17	19	Woman	Yes	5 to 9 years	1 to 3	2 to 4
P18	19	Woman	No†	10+ years	1 to 3	Irregular
P19	19	Man	No†	5 to 9 years	0.5 to 1	1
P20	26	Man	No	10+ years	1 to 3	5+
P21	21	Woman	Yes†	10+ years	1 to 3	2 to 4
P22	30	Queer	Yes†	10+ years	3+	5+
P23	22	Man	No	5 to 9 years	0.25 to 0.5	2 to 4
P24	23	Man	Yes	10+ years	1 to 3	5+
P25	22	Woman	No	10+ years	0.5 to 1	Irregular
P26	20	Undisclosed	Yes	10+ years	0.5 to 1	2 to 4
P27	37	Man	No	10+ years	3+	Irregular
P28	19	Man	Yes	5 to 9 years	1 to 3	1
P29	19	Woman	No	1 to 4 years	0.25 to 0.5	2 to 4
P30	20	Man	Yes	10+ years	0.5 to 1	2 to 4
P31	19	Man	Yes	10+ years	1 to 3	5+
P32	19	Woman	No	1 to 4 years	1 to 3	1
P33	21	Man	No	None	0.5 to 1	2 to 4
P34	20	Man	No†	5 to 9 years	0.5 to 1	2 to 4
P35	18	Woman	No	None	0.25 to 0.5	Irregular
P36	20	Man	Yes†	10+ years	0.5 to 1	2 to 4
P37	18	Man	Yes	10+ years	1 to 3	5+
P38	18	Man	Yes	5 to 9 years	1 to 3	2 to 4
P39	18	Man	Yes	10+ years	0.25 to 0.5	2 to 4
P40	20	Man	Yes	10+ years	1 to 3	5+
P41	18	Man	Yes	5 to 9 years	1 to 3	5+
P42	20	Woman	Yes	5 to 9 years	0.5 to 1	2 to 4
P43	18	Man	Yes	5 to 9 years	0.5 to 1	5+

†: indicates a more complicated response than Yes/No

Table H.10 (continued)

	Demographics		Gaming Information			
ID	**Age**	**Gender**	**Gamer**	**History**	**Length (hours)**	**Sessions per week**
P44	18	Man	Yes	10+ years	0.5 to 1	5+
P45	20	Woman	No	< 1 year	0.25 to 0.5	5+
P46	18	Man	Yes	10+ years	1 to 3	2 to 4
P47	20	Woman	Yes	5 to 9 years	0.5 to 1	2 to 4
P48	21	Man	Yes	10+ years	0.5 to 1	2 to 4
P49	23	Woman	Yes	5 to 9 years	1 to 3	2 to 4
P50	19	Woman	Yes	10+ years	1 to 3	5+
P51	19	Man	Yes	10+ years	1 to 3	5+
P52	18	Man	Yes	10+ years	0.5 to 1	5+
P53	20	Man	No	5 to 9 years	Irregular	Irregular
P54	21	Man	No	5 to 9 years	0.5 to 1	1
P55	21	Man	Yes	10+ years	1 to 3	5+
P56	19	Man	No	None	0.25 to 0.5	Irregular
P57	20	Man	No†	10+ years	0.5 to 1	2 to 4
P58	26	Woman	No	10+ years	1 to 3	1
P59	28	Woman	Yes	10+ years	1 to 3	2 to 4
P60	21	Woman	Yes	10+ years	1 to 3	5+
P61	33	Woman	No	None	0.25 to 0.5	Irregular
P62	24	Woman	Yes	10+ years	1 to 3	5+
P63	19	Woman	No	5 to 9 years	0.5 to 1	5+
P64	18	Trans Masc/ Non-binary	Yes	10+ years	0.25 to 0.5	2 to 4
P65	23	Woman	Yes	5 to 9 years	1 to 3	2 to 4
P66	25	Woman	No	1 to 4 years	0.25 to 0.5	Irregular
P67	20	Woman	No	1 to 4 years	0.5 to 1	5+
P68	26	Woman	No	None	0.25 to 0.5	Irregular
P69	31	Man	Yes	10+ years	3+	5+
P70	27	Man	Yes	10+ years	3+	5+
P71	27	Man	Yes	10+ years	1 to 3	5+

†: indicates a more complicated response than Yes/No

Table H.10 (continued)

	Demographics		Gaming Information			
ID	**Age**	**Gender**	**Gamer**	**History**	**Length (hours)**	**Sessions per week**
P72	27	Man	Yes	10+ years	1 to 3	5+
P73	29	Man	Yes	10+ yeas	1 to 3	5+
P74	20	Man	Yes	10+ years	3+	5+
P75	33	Woman	Yes	10+ years	1 to 3	2 to 4

†: indicates a more complicated response than Yes/No

H.3.2 Participant Group Distribution

Group	Participants
Easy	P03, P06, P09, P12, P15, P18, P20, P26, P29, P32, P35, P37, P40, P43, P46, P49, P52, P56, P59, P62, P69, P72, P75
Control	P02, P05, P08, P11, P14, P17, P19, P22, P24, P25, P28, P31, P34, P39, P42, P45, P48, P51, P54, P57, P60, P61, P64, P67, P70, P73
Hard	P01, P04, P07, P10, P13, P16, P21, P23, P27, P30, P33, P36, P38, P41, P44, P47, P50, P53, P55, P58, P63, P66, P68, P71, P74

Table H.11: Participant Groups

H.3.3 Participant Gaming Context

Table H.12 summarizes the participants' gaming platforms and a sample of the games they report.

Table H.12: Participant Gaming Context.

ID	PC	Ph/T	Con	HCon.	Sample of Reported Games/Genres
P01	✓	×	✓	×	LoL, Zelda, CoD, Stardew Valley, Mario Party
P02	✓	✓	✓	×	FIFA, CoD, Fortnite
P03	×	×	✓	×	Last of Us, GTA V, God of War
P04	✓	✓	✓	×	Rocket League, FIFA, Valorant
P05	✓	×	×	×	Valorant, Minecraft, Elden Ring, Undertale
P06	✓	✓	✓	✓	FPS, Minecraft, Action-Adventures, RPGs
P07	✓	×	✓	×	Rocket League, Rainbox Six Siege, NHL, FIFA
P08	✓	×	✓	×	Celeste, Inscryption, Zelda: Breath of the Wild, Hades, Pokemon
P09	✓	✓	✓	×	CoD, FIFA, RPG
P10	✓	✓	×	×	LoL, Tetris, Valorant, Roblox, Minecraft, Honkai: Star Rail, Maplestory
P11	✓	×	×	×	Assassin's Creed, Far Cry, Ori and the Blind Forest
P12	✓	✓	✓	✓	Dark Souls, Pokemon, PUBG, Osu!, Platformers, Smash Bros.
P13	✓	✓	×	✓	Valorant, Forza, FIFA, Assassin's Creed
P14	✓	✓	×	×	Super Mario Bros, Wii Sports, other Wii games
P15	✓	×	×	×	Genshin Impact, Fortnite
P16	✓	✓	×	×	GTA V, PUBG Mobile
P17	✓	✓	×	×	LoL, Minecraft, Genshin Impact
P18	×	×	✓	×	CoD

Legend: Computer (PC), Smartphone/Tablet (Ph/T), Console (Con), Handheld consoles (HCon), League of Legends (LoL), Call of Duty (CoD), Grand Theft Auto (GTA), Counterstrike: Global Offensive (CS:GO), Bloons Tower Defense (BTD), Legend of Zelda (Zelda), Final Fantasy (FF), First Person Shooter genre (FPS), Roleplaying Game genre (RPG), Massively Multiplayer Online (MMO), Multiplayer Online Battle Arena genre (MOBA), Play Unknown Battlegrounds (PUBG), Civilization (Civ), Magic the Gathering (MtG)

Table H.12 (continued)

ID	PC	Ph/T	Con	HCon.	Sample of Reported Games/Genres
P19	✓	×	✓	✓	Minecraft, Kingdom Hearts, Forza Horizon, CS:GO
P20	×	×	✓	×	FIFA, Overwatch, Warzone, Spiderman
P21	×	✓	✓	×	Animal Crossing, Mario Kart, Mario Party, Overcooked 2, Cookie Run Kingdom
P22	✓	×	✓	✓	MMORPGs, MOBAs, party games, fighting games, Mario Kart
P23	✓	✓	×	×	Clash of Clans, Hearthstone, LoL, Civ
P24	✓	✓	✓	×	Fortnite, Baldur's Gate 3, Mario Kart 8, Smash Bros. Ultimate
P25	✓	✓	×	×	Catan Online, Minecraft, Mario Kart, Animal Crossing
P26	✓	✓	×	×	Strategy games, RPGs
P27	✓	×	×	×	Uncharted Waters, City building, World of Tanks
P28	✓	✓	✓	×	Valorant, FIFA, Gran Turismo, Minecraft, Clash Royale, Clash of Clans, Geometry Dash
P29	×	✓	✓	×	–
P30	✓	×	×	✓	Celeste, Hollow Knight, Zelda
P31	✓	×	✓	×	Apex Legends, Zelda, indie games
P32	✓	×	✓	×	Nier Replicant, Omori, Steins Gate, Smash Ultimate
P33	×	✓	✓	×	FPS games, GTA, Racing games
P34	×	✓	✓	×	Rocket League, For Honour, God of War
P35	×	✓	×	×	Love Nikki, Rusty Lake
P36	✓	✓	✓	✓	Professor Layton, Zelda: Tears of the Kingdom, Skyrim, 2D platformers
P37	✓	×	×	×	Last of Us, Bioshock, Lethal Company, Fortnite

Legend: Computer (PC), Smartphone/Tablet (Ph/T), Console (Con), Handheld consoles (HCon), League of Legends (LoL), Call of Duty (CoD), Grand Theft Auto (GTA), Counterstrike: Global Offensive (CS:GO), Bloons Tower Defense (BTD), Legend of Zelda (Zelda), Final Fantasy (FF), First Person Shooter genre (FPS), Roleplaying Game genre (RPG), Massively Multiplayer Online (MMO), Multiplayer Online Battle Arena genre (MOBA), Play Unknown Battlegrounds (PUBG), Civilization (Civ), Magic the Gathering (MtG)

Table H.12 (continued)

ID	PC	Ph/T	Con	HCon.	Sample of Reported Games/Genres
P38	✓	×	×	×	Minecraft, Binding of Isaac, CS:GO, BTD6, Elden Ring, Lethal Company
P39	✓	✓	×	×	Minecraft, Open World Exploration, Co-Op games, **not** shooters
P40	✓	×	×	×	Minecraft, Europa Universalis IV, Rainbow Six Seige
P41	✓	×	×	×	FIFA, NBA, Valorant
P42	✓	✓	×	×	Minecraft, Stardew Valley, Team Fight Tactics, Persona, Mario Kart, Platformers
P43	✓	✓	×	×	Valorant, Clash of Clans, FIFA Mobile
P44	✓	✓	×	×	FIFA, Valorant, Spiderman
P45	✓	✓	×	×	Subway Surfers, GTA Vice City, Angry Birds, Racing and Adventure games
P46	✓	×	✓	×	Valorant, Spiderman, Forza, Gang Beasts
P47	✓	×	×	×	Valorant, Rimworld, CivVI, Skyrim
P48	✓	✓	×	×	LoL, Into the Breach, Slay the Spire
P49	✓	×	×	×	Binding of Isaac: Rebirth
P50	✓	✓	✓	×	FF XIV, Metal Gear Rising: Revengeance, Baldur's Gate 3
P51	✓	✓	×	×	Osu!, Overwatch 2, Genshin Impact
P52	✓	✓	×	×	FIFA, Fortnite, NBA
P53	✓	✓	×	×	Puzzle games (Mini Metro) and Thinking games (Hitman)
P54	✓	×	✓	×	Pokemon, Zelda: Tears of the Kingdom, Portal 2
P55	✓	✓	×	×	Genshin Impact, Metroidvanias, Portal, Valorant
P56	×	✓	✓	×	Fortnite, FIFA

Legend	Computer (PC), Smartphone/Tablet (Ph/T), Console (Con), Handheld consoles (HCon), League of Legends (LoL), Call of Duty (CoD), Grand Theft Auto (GTA), Counterstrike: Global Offensive (CS:GO), Bloons Tower Defense (BTD), Legend of Zelda (Zelda), Final Fantasy (FF), First Person Shooter genre (FPS), Roleplaying Game genre (RPG), Massively Multiplayer Online (MMO), Multiplayer Online Battle Arena genre (MOBA), Play Unknown Battlegrounds (PUBG), Civilization (Civ), Magic the Gathering (MtG)

Table H.12 (continued)

ID	PC	Ph/T	Con	HCon.	Sample of Reported Games/Genres
P57	✓	×	×	×	LoL, Valorant, Hearthstone, Pokemon, Mario
P58	✓	×	×	×	Genshin Impact, Don't Starve, MtG: Arena
P59	×	×	✓	×	CoD, Zelda, Overwatch!, Shadow of the Colossus
P60	✓	✓	×	×	Risk of Rain, Honkai: Star Rail, Hollow Knight, Maplestory
P61	✓	×	✓	×	Mortal Kombat, Crash Bandicoot, The Neverhood, GTA 3
P62	✓	✓	✓	×	CivV, Minecraft, Duolingo, Chess, Stardew Valley
P63	✓	×	×	×	Nier, Genshin Impact, LoL
P64	✓	✓	✓	✓	Fortnite, Minecraft, Sims, Assassin's Creed, Mario Kart
P65	✓	×	×	×	Puzzle games, FPS, Multiplayer Cooperation
P66	×	✓	×	×	Candy Crush, 2048
P67	✓	✓	×	×	Valorant, Osu!, Honkai: Star Rail, Bayonetta
P68	✓	✓	×	×	–
P69	✓	×	✓	×	Zelda, PUBG, LoL, Raft, Civ, Starcraft
P70	✓	×	✓	×	Minecraft, Zelda, No Man's Sky, Old School Runescape, CoD, Pac-Man, Vampire Survivors
P71	✓	✓	×	×	MOBAs, 4X, Strategy style
P72	✓	✓	×	×	LoL, Runescape, Jackbox, Idle games
P73	✓	✓	×	✓	Palworld, LoL, Pokemon
P74	✓	✓	✓	✓	Action-Adventure, JRPGs, Puzzle, Shooters, Platformers, Co-op, Arcade

Legend: Computer (PC), Smartphone/Tablet (Ph/T), Console (Con), Handheld consoles (HCon), League of Legends (LoL), Call of Duty (CoD), Grand Theft Auto (GTA), Counterstrike: Global Offensive (CS:GO), Bloons Tower Defense (BTD), Legend of Zelda (Zelda), Final Fantasy (FF), First Person Shooter genre (FPS), Roleplaying Game genre (RPG), Massively Multiplayer Online (MMO), Multiplayer Online Battle Arena genre (MOBA), Play Unknown Battlegrounds (PUBG), Civilization (Civ), Magic the Gathering (MtG)

Table H.12 (continued)

ID	PC	Ph/T	Con	HCon.	Sample of Reported Games/Genres
P75	✓	×	✓	×	Mario Kart, FF XII, Layers of Fear, Love Nikki, Pokemon Snap, Shadow of the Colossus

Legend: Computer (PC), Smartphone/Tablet (Ph/T), Console (Con), Handheld consoles (HCon), League of Legends (LoL), Call of Duty (CoD), Grand Theft Auto (GTA), Counterstrike: Global Offensive (CS:GO), Bloons Tower Defense (BTD), Legend of Zelda (Zelda), Final Fantasy (FF), First Person Shooter genre (FPS), Roleplaying Game genre (RPG), Massively Multiplayer Online (MMO), Multiplayer Online Battle Arena genre (MOBA), Play Unknown Battlegrounds (PUBG), Civilization (Civ), Magic the Gathering (MtG)

H.4 PXI Results

The following tables are the resulting statistical analyses for the PXI data.

H.4.1 Descriptive Statistics

Construct	Condition	Range	Mode	Median	Mean	Std. Dev
Meaningful	Easy	-6, 8	5	4	3.13	3.46
	Control	-9, 9	-6	0	0.19	5.01
	Hard	-7, 7	2	0	-0.92	3.67
Mastery	Easy	-9, 9	9	6	3.54	5.78
	Control	-7, 6	3	2	0.50	3.72
	Hard	-9, 7	1	0	-1.56	4.82
Immersion	Easy	-3, 9	6	6	4.42	3.61
	Control	-6, 9	6	4	3.35	4.04
	Hard	-9, 9	6	3	2.56	5.11
Progress	Easy	-8, 9	5	5	4.33	4.41
	Control	-6, 9	-5	1	1.50	5.26
	Hard	-9, 9	-4	3	1.48	5.52
Audiovisual	Easy	-1, 9	9	8	6.04	3.43
	Control	-4, 9	9	5	4.31	3.81
	Hard	-6, 9	4	4	4.32	3.78
Challenge	Easy	-7, 9	6	5	3.63	4.52
	Control	-8, 9	0	4	2.62	3.96
	Hard	-8, 7	5	0	0.56	4.86
Controls	Easy	5, 9	9	8	7.88	1.23
	Control	0, 9	9	8	6.69	2.53
	Hard	-2, 9	9	8	6.48	2.99
Goals	Easy	5, 9	9	9	8.25	1.23
	Control	5, 9	9	9	8.04	1.40
	Hard	0, 9	9	9	7.56	2.27

Table H.13: Descriptive statistics of Construct scores. Construct scores can range from -9 to 9.

H.4.2 Checking Group Differences

Construct	Sum Squares	Mean Square	F(2,74)	Sig.
Meaningful	**213.18**	**106.59**	**6.26**	**0.003**
Mastery	**321.97**	**160.98**	**6.93**	**0.002**
Immersion	42.47	21.23	1.15	0.324
Progress	131.93	65.96	2.67	0.076
Audiovisual	48.73	24.37	1.76	0.174
Challenge	*120.25*	*60.12*	*3.03*	*0.054*
Controls	27.60	13.80	2.43	0.095
Goals	6.17	3.08	1.07	0.349

Table H.14: Between group results for One-way ANOVAs of Construct scores. Significant constructs are bolded. Constructs requiring more investigation are italicized.

Table H.15: Tukey HSD Results for PXI Constructs. Significant pairs are bolded.

Construct	Groups	Mean Diff	Std. Error	Sig.
	AB	**2.93**	**1.17**	**0.038**
Meaningful	**AC**	**4.05**	**1.18**	**0.003**
	BC	1.11	1.16	0.603
	AB	3.04	1.36	0.073
Mastery	**AC**	**5.10**	**1.38**	**0.001**
	BC	2.06	1.35	0.285
	AB	1.07	1.22	0.655
Immersion	AC	1.86	1.23	0.293
	BC	0.79	1.20	0.792
	AB	2.83	1.41	0.116
Progress	AC	2.85	1.42	0.117
	BC	0.02	1.39	1
	AB	1.73	1.04	0.227
Audiovisual	AC	1.72	1.05	0.238
	BC	-0.12	1.03	1
	AB	1.01	1.26	0.704
Challenge	**AC**	**3.07**	**1.27**	**0.048**
Legend	*A = Easy, B = Control, C = Hard.*			
	Mean Diffs are first group minus second group (e.g. AB is A-B)			

Table H.15: Tukey HSD Results for PXI Constructs. Significant pairs are bolded.

Construct	Groups	Mean Diff	Std. Error	Sig.
Controls	BC	2.06	1.25	0.233
	AB	1.18	0.67	0.192
	AC	1.40	0.68	0.108
	BC	0.21	0.67	0.946
Goals	AB	0.21	0.48	0.899
	AC	0.69	0.49	0.335
	BC	0.48	0.48	0.576
Legend	*A = Easy, B = Control, C = Hard.*			
	Mean Diffs are first group minus second group (e.g. AB is A-B)			

H.4.3 Contextualizing with Individual Items

Table H.16: PXI Descriptive Statistics per Item. Item scores can range from -3 to 3.

	Group	Range	Mode	Median		Group	Range	Mode	Median
Meaning					**Audiovisual Appeal**				
Mean1	Easy	-2, 3	2	1	AV1	Easy	-1, 3	3	3
	Control	-3, 3	0	0		Control	-2, 3	3	2
	Hard	-3, 2	0	-1		Hard	-2, 3	3	2
Mean2	Easy	-2, 3	2	1	AV2	Easy	-1, 3	3	3
	Control	-3, 3	0	0		Control	-2, 3	1	1
	Hard	-3, 2	-2	0		Hard	-2, 3	1	1
Mean3	Easy	-2, 3	2	2	AV3	Easy	-1, 3	3	2
	Control	-3, 3	0	0		Control	-2, 3	3	2
	Hard	-3, 3	-1	0		Hard	-2, 3	2	2
Mastery					**Challenge**				
Mast1	Easy	-3, 3	3	2	Cha1	Easy	-3, 3	2	2
	Control	-2, 2	1	0		Control	-3, 3	2	1
	Hard	-3, 2	1	0		Hard	-3, 3	2	1
Mast2	Easy	-3, 3	3	2	Cha2	Easy	-3, 3	2	2
	Control	-2, 3	1	1		Control	-3, 3	2	1

Table H.16: PXI Descriptive Statistics per Item. Item scores can range from -3 to 3.

	Group	Range	Mode	Median		Group	Range	Mode	Median
	Hard	-3, 2	1	1		Hard	-2, 2	2	1
Mast3	Easy	-3, 3	3	2	Cha3	Easy	-3, 3	2	2
	Control	-3, 2	-1	-1		Control	-2, 3	1	1
	Hard	-3, 3	-3	-2		Hard	-3, 3	-2	0
Immersion					**Ease of Control**				
Imm1	Easy	-2, 3	1	1	Con1	Easy	1, 3	3	3
	Control	-3, 3	2	1		Control	0, 3	3	3
	Hard	-3, 3	2	1		Hard	-2, 3	3	3
Imm2	Easy	-2, 3	1	1	Con2	Easy	1, 3	3	3
	Control	-3, 3	2	2		Control	0, 3	3	3
	Hard	-3, 3	2	2		Hard	-2, 3	3	3
Imm3	Easy	-1, 3	3	2	Con3	Easy	1, 3	3	3
	Control	-3, 3	2	2		Control	0, 3	3	2
	Hard	-3, 3	2	2		Hard	-2, 3	3	2
Progress Feedback					**Goals and Rules**				
Prog1	Easy	-3, 3	3	2	Goal1	Easy	2, 3	3	3
	Control	-3, 3	3	1		Control	1, 3	3	3
	Hard	-3, 3	2	1		Hard	-2, 3	3	3
Prog2	Easy	-2, 3	1	2	Goal2	Easy	2, 3	3	3
	Control	-3, 3	2	1		Control	1, 3	3	3
	Hard	-3, 3	2	1		Hard	1, 3	3	3
Prog3	Easy	-3, 3	1	1	Goal3	Easy	1, 3	3	3
	Control	-3, 3	3	0		Control	2, 3	3	3
	Hard	-3, 3	2	1		Hard	1, 3	3	3

H.5 Coding

The following is a summary of codes from the thematic analysis data. We provide an excerpt example for each code to illustrate how it manifests. In cases where interpretation is less clear-cut, we provide multiple examples to build a sense of what we see.

Table H.17: List of codes generated from data.

Code	**Shorthand**	**Explanation**	**Example(s)**
Game was easy for me	Easy	A lack of felt/perceived difficulty	"Difficulty was not an overall factor..." (P69)
Approaching the game with a strategy	Strategy	Discussions about optimal and suboptimal ways of approaching the game	"...ensuring position of my hands..." (P69) "...I should have spent more time figuring out what techniques worked best..." (P58)
Understanding input	Input	The description of the physical actions/interactions with the input device	"...rapidly click the right arrow key..." (P69)
Understanding of mechanics	Mechanic	Descriptions and discussion of the basic instructions, rules, and mechanics of the game.	"...rapidly click the right arrow key..." (P69) "...the game was fairly simple." (P19)
Fatigue	Fatigue	Feelings of being tired and the impact it has on performance	"...without succumbing to fatigue" (P69)
Speed and pace as important factors in performance	Pace	Discussion of speed, pace, and consistency of input as it relates to game performance.	"starting off fast and then getting slower as the timer went on so keeping the pace was the hardest part" (P46)

Table H.17 (continued)

Code	Shorthand	Explanation	Example(s)
Feeling unsure about the effectiveness of your approach	Uncertainty	Discussions about the player's perception of their performance in relation to game factors.	"...so I did not know how my technique was working..." (P51) "...I didn't know what type of button pressing would work best..." (P20)
Technological limitations as an important factor in performance	Technology	Discussion about the specifics of interfaces and controls in relation to performance and/or response recognition.	"The buttons on this keyboard are much less tactile than the one I'm used to..." (P63) "...being on the membrane keyboard was valid but I would love to try again with my own mechanical keyboard." (P70) "...wished I had a mechanical keyboard for some haptic feedback on the button presses to know when I could begin a new press." (P71)
Game did not tell me how well I was doing in real-time	Feedback	Discussion of response recognition from the game-end through visual or audio means.	"...difficult to tell what counted as one tap..." (P35) "feedback DURING the game would vastly improve performance" (P69)

Table H.17 (continued)

Code	Shorthand	Explanation	Example(s)
This was enjoyable	Positive Experience	General statements about enjoying themselves, having a good time, etc.	"...I enjoyed all the games and they were well made and I had a lot of fun with it..." (P05)
I am better than this performance lets on	Expectations	Emotionally-charged discussions of rank/performance that imply they believe they were helped or hindered by the game in some way.	"...I do feel like I was in the hard rank..." (P70) "...results at time felt a bit arbitrary." (P42) "...but I don't usually do well with rhythm games" (P58)
I need to be motivated to perform	Confidence	Discussions about the player's perception of their own expertise and performance in relation to ways the game could improve.	"...Getting a B rank did demotivate me in the game."(P70)
Focusing on the game	Attention	Specifically discussing attention, or task switching as a limiting factor or difficult part of the game.	"...I would trip up and forget to alternate between the keys and cost myself presses." (P24) "...ensuring I had a proper rhythm..." (P54) "Fire starting game needed a lot of attention..." (P70)

Table H.17 (continued)

Code	Shorthand	Explanation	Example(s)
The biomechanics of the game	Ability	References to physical limitations.	"...my two hands to fall out of sync..." (P28) "Just controlling my arm again. You have more muscles active two have the two fingers..." (P57) "Holding hand in a more awkward position to reach both buttons" (P25)
I am/am not this type of gamer	Experience	Discussions about the player's gaming literacy and history.	"...I'm not much of a button masher..." (P01) "...but I don't usually do well with rhythm games" (P58)
Modulating effort based on perceived performance	Modulating	The ways that effort are conserved or applied based on how they are doing	"...I don't feel motivated to button mash as hard as possible unless it's in a competitive setting, or a reward I genuinely want." (P58)
Emotional regulation in response to performance	Emotional Regulation	The emotional response to performance and ways it is communicated	"Not getting angry at a bad rating" (P74) "...did not enjoy the taste of failure..." (P58)

Table H.17 (continued)

Code	Shorthand	Explanation	Example(s)
I had to think about what I was doing	Overload	When the difficulty resulted in an unanticipated amount of effort	"...I had to think about what I was doing. This slowed me down..." (P30)
Conceptual model of gameplay	Conceptualization		"...ensuring position of my hands to rapidly click the right arrow key..." (P69) "...button mashing games are usually pretty easy and don't have high thresholds for success..." (P58) "...the game was fairly simple." (P19)
Justifying their performance through meta-game knowledge	Justifying performance		"Difficulty was not an overall factor, but ensuring position of my hands..." (P69) "...I should have spent more time figuring out what techniques worked best, but I wasn't expecting to need to optimize it that much because button mashing games are usually pretty easy and don't have high thresholds for success..." (P58)