TEACHING FOR TRANSFER

TEACHING FOR TRANSFER:
A RETRIEVAL-BASED INTERVENTION, AND
A PUTATIVE TOOL TO GAUGE LEARNING APPROACHES

By ANDREW B. LOGIUDICE, B.Sc. (Honours)

A Thesis Submitted to the School of Graduate Studies in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

DOCTOR OF PHILOSOPHY (2020)	Psychology, Neuroscience & Behaviour
McMaster University	Hamilton, Ontario
TITLE:	Teaching for transfer: A retrieval-based intervention, and a putative tool to gauge learning approaches
AUTHOR:	Andrew B. LoGiudice, B.Sc., McMaster University
SUPERVISOR:	Sandra Monteiro, PhD
NUMBER OF PAGES:	xv, 124

Lay Abstract

A central goal of education is to equip students with ‘flexible’ knowledge, enabling them to *transfer* the far-reaching principles they have learned to solve new, real-world problems. But what conditions of training are most conducive to transfer? One understudied technique involves being tested on the same principle in dissimilar contexts. The experiments reported in Chapter 2 provide evidence for this training technique in the domain of problem solving. Aside from direct interventions, another approach has been to measure individual differences among students to predict how much they engage in “deep learning”—a process closely associated with transfer. However, four correlational studies in Chapters 3 and 4 revealed little support for this approach, highlighting the difficulty of characterizing learning strategies using self-reports. In sum, this shows promise for interventions involving repeated testing in dissimilar contexts, but little promise for a self-report inventory meant to capture individual differences in student learning.

Abstract

The phenomenon of *transfer*—our ability to perform novel tasks by generalizing from past experiences—has long captivated theorists and practitioners. As educators it is essential for us to understand what types of learning best promote transfer and to structure our curricula accordingly. With that goal in mind, this dissertation outlines two lines of research.

For the first line of research I adopted an experimental approach in the domain of problem solving, examining a training technique whereby the learner solves practice problems for the same principle in dissimilar contexts as opposed to highly similar contexts. The key finding was that contextual variability improved transfer outcomes when a set of training problems were solved spaced in time (akin to a closed-book test), but not when prior training problems and their solutions remained visible throughout training (akin to an open-book test). This finding suggests that contextual variability during training can be beneficial because it forces the learner to more effortfully recall what they have learned in the past.

For the second line of research I then adopted a correlational approach, investigating a ubiquitous self-report inventory, the Study Process Questionnaire (SPQ), which is meant to quantify student learning approaches to predict educational outcomes. However, the SPQ's predictive validity has recently been challenged because deep learning and its corresponding outcomes remain poorly defined. To tackle this measurement issue, my colleagues and I operationally defined outcome measures in real university courses to tap more precisely into transfer of learning. Across several studies we found limited evidence for the SPQ's ability to predict transfer outcomes, leading us to suggest that educators and researchers should be more cautious about using this self-report inventory to characterize student learning.

Acknowledgements

Well, the end looks a lot different than what I expected. Science is certainly a humbling teacher. But for all the hardships, here I am. I'm grateful that this experience has forced me to look more deeply at myself, what I value, and the things that makes me tick. Above all else, the past few years have strengthened my love of writing, my appreciation of scientific curiosity in all forms, and also my desire to be compassionate to others in times of uncertainty or despair. And if that isn't nice, then what is?

Geoff Norman, after five dizzying years, the one thing I can say for certain is that I made it this far because of you. Your relentless passion for science and storytelling have inspired me to keep going. Thank you for reminding me to take a step back from all the work and appreciate that science is a deeply human endeavour, because it's well-rounded people that make science so great. P.S. You've bought me enough beer. For the love of all that is good, the next is on me.

Sandra Monteiro, it feels like we've worked together for far longer than we actually have, which is a testament to how positive our interactions have been. I look forward to our meetings because I know we can have an in-depth conversation while still having some fun and sharing a good laugh. Thank you for your compassion, words of encouragement, insights into the strange but rewarding realm of medical education, and for reminding me to be proud of my work.

Bruce Milliken, we've never directly worked together on a project, which is a shame, but few people have taught me more about the nuts and bolts of science than you. I've listened carefully to your advice and it has greatly improved my work. I'll look back fondly on your contagious enthusiasm, thoughtful feedback, and appreciation of elegant, no-nonsense experiments. Academia needs more mentors like you.

David Shore, you taught me the joy of good writing, a lesson I'll never forget. The writing in this thesis would not have been as coherent without the many hours I spent as a student and teaching assistant for your writing course. Thank you for also embracing the personal side of research. At times I was made to feel like an outsider, but you were always the first to extend an invitation and offer words of advice. It meant the world.

Joe Kim, you kept my passion for teaching alive despite my ongoing frustrations with research. I would not be as strong a student today without your many lessons on storytelling and

effective communication. Here's hoping that our mutual passion for teaching keeps us in touch long after my graduation.

To my friends and colleagues in the department, especially the Millikenites, Shorians, and members of CogSci: this journey would not have taken shape without you. I was not always the calibre of friend or student that I aspire to be, and for that I'm regretful. But I will cherish the many memories of coffee runs, hallway shenanigans, the official summer leisure reading group (NOT a book club), and the constant words of support. Grad school wouldn't be the same without these small moments of kindness.

Lastly, to my friends and family, words cannot express what your support and patience have meant to me during these challenging years. Anthony, Brett, Harry, Kurt, Mac, Mike[2], Mitch, I view you all as brothers. Thank for you keeping me sane, humble, and laughing amidst the chaos. Lucia, Natalie, Kate, Anna, and especially Laura, your kindness and strength are unparalleled. Thank you for showing me that emotion and vulnerability are not weaknesses. Mom and Dad, my god it feels like forever since I started university. I know it wasn't easy to accept that I'd be away and busy (read: barely making money) for so long. I am forever indebted to you for encouraging me to explore my passions regardless of the uncertainty involved. You are the best parents a son could ask for.

Ending on the cheesiest note possible, I couldn't resist including lyrics from a song that stuck with me as I buckled down to finish this thesis. The song is *Moon of Day* by a little-known band called Attack in Black from Welland, Ontario. To me these lyrics are a reminder of everything we strive for as scientists. Mother Nature is out there, right now, doing what she does best with a trillion different moving parts. It ain't easy, and it ain't always glamorous, but I'll be damned if trying to piece together those parts isn't one of the most rewarding jobs in the world.

Have you written choruses
of movements earthly with no sound?
But inside your soft, orchestral art
rhythms of the world are found

Have you only to begin
to fall at morning's feet?
To hold at once what word was mine
and in your absence keep

Moon of day

Table of Contents

List of Figures

Chapter 2

Chapter 3

Chapter 4

List of Tables

Chapter 3

List of Abbreviations and Symbols

SPQ: Study Process Questionnaire

Declaration of Academic Achievement

This sandwich thesis contains three manuscripts. Chapter 4 has been submitted to scholarly journals for peer review. Chapters 2 and 3 will also be submitted for peer review once my co-authors and I have decided on the appropriate journals. I have obtained permission from all coauthors to include these manuscripts in my dissertation. If accepted for publication, permission from the respective copyright holders to reprint the manuscripts will also be obtained. For all these studies I conceptualized the designs, conducted literature reviews, collected data myself or oversaw data collection, ran all statistical analyses, and prepared the manuscripts.

Because of the sandwich format of the thesis, the three data chapters (2, 3, and 4) are meant to be read as standalone documents, and so there is considerable redundancy, most notably in Chapters 3 and 4. Figures, tables, and references are included at the end of each data chapter in the form appropriate for submission to journals. The roles of co-authors, when the studies were conducted, and relevant conference presentations for each data chapter are outlined below.

Chapter 2

> **LoGiudice, A. B.**, Norman, G. R., Monteiro, S., Kulasegaram, K. M., & Watter, S. (in preparation). A synergistic effect of contextual variability and spacing on subsequent transfer.

Geoff Norman provided support with the conceptualization of the study, scoring, data analysis, interpretation of the findings, and manuscript preparation. Mahan Kulasegaram provided support with the development of learning materials, scoring, and manuscript preparation. Sandra Monteiro assisted with data analysis, interpretation of findings and manuscript preparation. Scott Watter provided support with resources and data collection. The data were collected during the 2018–2019 academic terms at McMaster University. I presented these data as a poster at the Annual Meeting of the Psychonomic Society in 2019.

Chapter 3

> **LoGiudice, A. B.**, Saadat, P., Vale, J., Clemmer, R., Gordon, K., Ahmad, A., Barrington, J., Cassidy, R., Monteiro, S., and Norman, G. R. (in preparation). Do self-reported learning approaches predict transfer, and just how context-specific are they?

Pakeezah Saadat assisted with the conceptualization of the studies, data collection, and manuscript preparation. Julie Vale, Ryan Clemmer, and Karen Gordon assisted with the conceptualization and data collection for Study 2 as well as preparation of the manuscript. Arshad Ahmad, Janette Barrington, and Robert Cassidy assisted with the conceptualization of Study 1 and manuscript preparation. Sandra Monteiro and Geoff Norman assisted with the conceptualization of the designs, data analysis, interpretation of the data, and manuscript preparation. The data from Study 1 were collected during the 2018–2019 academic terms at McMaster University. The data from Study 2 were collected during the 2019–2020 academic terms at both McMaster University and the University of Guelph. I gave an oral presentation on these data at the McMaster Conference on Education & Cognition in 2018.

Chapter 4

> **LoGiudice, A. B.**, Norman, G. R., Manzoor, S., & Monteiro, S. (submitted). Do deep learning approaches predict transfer of knowledge? Operationally defining transfer outcomes in practice via student perceptions.

All authors helped conceptualize the design of the studies. Geoff Norman and Sandra Monteiro assisted with data analysis, data interpretation, and manuscript preparation. Saba Manzoor assisted with data collection and preparation of the final manuscript. The data from both studies were collected during the Fall 2020 academnic term at McMaster University. These data have not yet been presented at any conferences.

This manuscript was submitted to the Journal of Educational Psychology but was turned down because this journal does not focus on validation studies. I have since re-submitted it to the journal Higher Education.

Chapter 1: General Introduction

> "It is the glory of geometry that from so few principles, fetched from without, it is able to accomplish so much."
>
> —Sir Isaac Newton, 1729

Transfer: A Bird's Eye View

Why do we teach? All educators know the well-intentioned student who memorizes every fact and superficial detail. But we do not teach topics like physics just so students can recall relevant facts and details; we teach physics so they can understand and apply larger principles like the laws of thermodynamics, electromagnetism, or gravitation. Likewise, we do not teach principles like gravity just so students can predict how long an object will take to fall; we teach gravity so they understand *why* objects fall and how the same principle is remarkable because it generalizes to explain a host of other phenomena. So, by teaching an assortment of principles across domains, our hope is that students will gain some flexibility in their knowledge, allowing them to *transfer* their understanding of far-reaching principles to new and unfamiliar problems (see Bransford & Schwartz, 1999; Halpern & Hakel, 2003; Salomon & Perkins, 1989).

Unsurprisingly, ideas about transfer are central to many fields of study, including but not limited to motor learning (e.g., Jarus & Goverover, 1999; Landin et al., 1993; Newell & Shapiro, 1976), category learning (e.g., Brooks & Vokey, 1991; Jacoby et al., 2010; Kornell & Bjork, 2008; Whittlesea & Dorken, 1993), memory (e.g., Butler, 2010; Carpenter, 2012; Morris et al., 1977), problem solving (e.g., Holyoak & Koh, 1987; Needham & Begg, 1991; Ross, 1987; for review, see Reeves & Weisberg, 1994), and education more broadly (e.g., Billing, 2007; Day & Goldstone, 2012; Norman, 2009; Salomon & Perkins, 1989). Owing to this diversity, transfer and its underlying mechanisms are often described differently depending on one's field of study or the specific learning task under investigation. And yet the core questions remain largely the same: What constitutes transfer, when does it occur, and how?

Context Specificity

Early work sparked heated debates about what transfer is and how it ought to be studied. The dominant view in the early 20th century was that of formal discipline, which suggests that broad mental functions like memory and reasoning could be practised and improved *in general*

through repetitive tasks (e.g., Lewis, 1905). By this view, someone might improve their general memory capacity by memorizing and recalling poems or texts verbatim, or their general reasoning skills by repeatedly solving open-ended problems. However, Thorndike and Woodsworth (1901) challenged this view by showing that improvements on one task often did not transfer to similar tasks that ought to require the same mental function. So began discussions about transfer being rare and difficult to study because it is extremely context-specific[1], meaning that performance on a novel task depends on how closely it resembles the conditions of initial learning (also see Bassok & Holyoak, 1989; Bransford & Schwartz, 1999; Morris et al., 1977).

The phenomenon of context specificity has since been reinforced through basic attention and memory experiments. Laboratory studies consistently show that performance on a novel task depends on the task's similarity to prior training events in terms of one's physical environment (for review, see Smith & Vela, 2001), but more so in terms of the specific characteristics of stimuli (e.g., Brooks et al., 1991; Brooks & Hannah, 2006; Gobet & Simon, 1996; Vokey & Brooks, 1992) or task demands (e.g., Morris et al., 1977; Logan, 1996; Whittlesea & Dorken, 1993; Whittlesea et al., 1994). For example, revealing the importance of similarity in stimuli between training and test, Logan (1996) found faster reaction times and fewer errors on a word categorization task when the word's colour remained constant across training and test. However, this was only true when participants were forced to attend to the colour dimension at both training and test, revealing the impact of similarity in both stimuli *and* task demands. The learning of relatively simple operations thus seems to be highly context specific.

More complex problem-solving tasks also seem to be context specific, as shown in studies of *analogical transfer*—i.e., problem solving by analogy to prior examples. The main question is how learners extract the higher-order relations between elements within an example, thereby permitting transfer of its "deep structure" to new problems (e.g., Gentner et al., 2003; Gentner & Smith, 2012; Markman & Gentner, 1993). But problem solving success is also strongly driven by the extent to which a problem resembles prior examples in terms of "surface features", the particular (and often arbitrary) characteristics of an example or problem that

[1] I adopt the term context specificity for consistency and ease of exposition, but the same general idea is often described using terms like encoding specificity (e.g., Tulving & Thomson, 1973), content specificity (e.g., Eva et al., 1998), domain specificity (e.g., Bassok & Holyoak, 1989), context-dependent memory (e.g., Smith & Vela, 2001), and transfer-appropriate processing (e.g., Morris et al., 1977).

instantiate it in some concrete way (for reviews, see Day & Goldstone, 2012; Reeves & Weisberg, 1994). For example, revealing the impact of even slight contextual changes, Holyoak and Koh (1987) found that fewer participants solved Duncker's (1945) X-ray problem when only a few words were changed in the training examples to refer to waves rather than lasers. Similarly, Ross (1987) had participants solve mathematics problems using formulas and manipulated whether the entities in a new problem (e.g., *mechanics* fixing *cars*) aligned with either the same or different variables of the formula relative to a prior example. As expected, this "cross-mapping" of surface features between training and test impaired transfer (also see Gentner & Toupin, 1986; Ross, 1989). Such findings leave little doubt: any sensible theory of transfer must explain these robust context specificity effects.

Instance Models

Accordingly, transfer phenomena are often explained using instance models of memory that assign a pivotal role to context. These models assume that abstract and generalizable knowledge only emerges as a result of discrete, context-laden experiences—instances—being stored in memory over time (e.g., Brooks, 1978; Hintzman, 1984; Jamieson et al., 2012; Logan, 1988; Medin & Schaffer, 1978; Whittlesea, 1997). For example, the MINERVA model (Hintzman, 1984) posits that episodic traces are stored in memory and later activated in parallel based on their similarity to retrieval cues, thereby producing an "echo" that reflects the combined output of all activated traces. A similar logic can be seen Whittlesea's (1997) SCAPE model, according to which we construct experiences—including stimuli and any mental operations performed on them—that become preserved in memory. These experiences are then said to be recruited in memory based on their similarity to retrieval cues, guiding the construction of a new experience. So, subsuming the idea of context specificity, these instance models assume the similarity between the particular conditions of training and test are critical for determining whether transfer will occur.

To illustrate the 'scaling up' of instance theories to explain problem solving, consider a popular framework that distills transfer into three steps: (i) encoding the target problem; (ii) recruiting prior examples in memory based on their similarity to the target problem—either in terms of surface features or deep structure; and (iii) mapping the relations between elements from an example onto the corresponding elements of the target problem (see Day & Goldstone, 2012; Gentner & Smith, 2012; Reeves & Weisberg, 1994). Learners are thus said to solve new

problems by quickly recruiting similar prior examples in memory that guide their solution attempts. And because learners are often unaware that prior examples are influencing their behaviour, the recruitment of prior instances is partly automatic and unavailable to consciousness (e.g., Day & Gentner, 2007; Day & Goldstone, 2011; Kostic et al., 2010; Schunn & Dunbar, 1996). In sum, this framework for problem solving suggests that performance is driven by similarity in context between training and test, even if the learner is often unaware of it.

Teaching for Transfer: How Far Have We Come?

These theories of transfer raise some important practical questions about educational practice. How well do our modern educational systems achieve this lofty goal of helping students overcome context specificity, empowering them to transfer far-reaching principles to new contexts? Addressing this question in turn requires a more fundamental understanding of the mind and how it acquires new information. Put simply, what conditions of learning are most conducive to transfer, and why?

Studies on these matters can be divided into two broad approaches, mirroring Cronbach's (1957) distinction between two disciplines in psychology: the experimental approach and the correlational approach. The experimental approach involves the manipulation of variables during training to examine their effects on subsequent transfer, whereas the correlational approach involves the identification of variables pertaining to learners (e.g., stable individual differences) or the learning environment that predict whether transfer will occur. These approaches complement each other; while the experimental approach homes in on candidate variables that can be manipulated one-by-one in a controlled manner, the correlational approach embraces the natural variability between learners to identify factors most predictive of transfer.

In what follows, I summarize key findings from each of these approaches and highlight questions that are addressed throughout this thesis. The first section reflects the experimental approach, focusing on studies in cognitive psychology concerning training interventions that are known to enhance subsequent transfer. Data speaking to two of these interventions are then presented in Chapter 2. The second section reflects the correlational approach, focusing on measurement tools that have been designed to predict a student's tendency to learn such that they can later transfer their knowledge to new contexts. Data speaking to one of these measurement tools are presented in Chapters 3 and 4.

Experimental Approaches

Emphasizing Deep Structure

The most straightforward way to improve transfer outcomes is to emphasize the deep structure of examples during training. This can be accomplished in at least two ways. First, the relations between key elements of an example can be highlighted or explicitly labelled during training (e.g., Brown et al., 1986; Catrambone, 1996; Kotovsky & Gentner, 1996; Loewenstein & Gentner, 2005; Miyatsu et al., 2018). For instance, Brown et al. (1986) asked children to read a story and manipulated whether they received questions about the key elements (e.g., about the protagonist and challenges they faced). Simply asking them these questions increased their probability of solving a related new problem. Second, the surface features of examples can be removed (e.g., Garner et al., 1989; Mayer et al., 2008) or made less salient during training (e.g., Goldstone & Sakamoto, 2003; Kaminski et al., 2013; Markman & Gentner, 1993; also see Fyfe et al., 2014), which presumably enhances learning by helping the learner home in on the deep structure of the example. Such findings underscore the need to direct the learner's attention to the deep structure of examples as opposed to nonessential details.

Case Comparison

A similar way to improve transfer outcomes is to have participants compare two or more examples of the same principle (e.g., Catrambone & Holyoak, 1989; Christie & Gentner, 2010; Gick & Holyoak, 1983; Richland & McDonough, 2010; for meta-analytic review, see Alfieri et al., 2013). Identifying similarities among examples is thought to promote structural alignment, referring to the encoding of deep structure by mapping the higher-order relations between key elements of one example onto the corresponding elements of another example (for brief review, see Gentner & Smith, 2012). Interestingly, learners do not always spontaneously compare examples even when it is beneficial to do so, as evidenced by learners exhibiting better transfer after being explicitly told to identify similarities between examples than after receiving the same examples without such instructions (see Alfieri et al., 2013)[2]. Thus, transfer outcomes can

[2] Even in the absence of explicit instructions, there is still some evidence that learners will spontaneously engage in comparison if the format of training makes it clear that the examples belong to the same principle (e.g., Quilici & Mayer, 2002; Ross & Kennedy, 1990).

sometimes be enhanced by encouraging learners to compare multiple examples of the same principle, because learners do not always adopt this strategy spontaneously.

Retrieval Practice

Another promising means of enhancing transfer is through *retrieval practice*, a technique whereby previously studied information is repeatedly retrieved from memory during training (for reviews, see Kornell & Vaughn, 2016; Roediger & Butler, 2011). The main thrust of this literature is that incorporating retrieval opportunities during training often elicits greater retention of the content relative to study-only control conditions (for meta-analysis, see Rowland, 2014). Most critical for present purposes, retrieval practice is also known to enhance subsequent transfer (e.g., Butler, 2010; Butler et al., 2017; Eglington & Kang, 2018; Jacoby et al., 2010; Larsen et al., 2013; Rohrer et al., 2010; for review, see Carpenter, 2012).

Most evidence for improved transfer following retrieval practice comes from studies on relatively simple comprehension, memory, or categorization tasks (see Carpenter, 2012). For example, Butler (2010) found that retrieval practice of passages during training led to better performance on subsequent inference questions than did study-only control conditions (also see Butler et al., 2017; Chan et al., 2006). Similarly, Rohrer et al. (2010) reported that retrieval practice of individual region names on a map led to better transfer on a final test that required more detailed knowledge of the spatial relations between regions. Extending this finding to the categorization of natural concepts, Jacoby et al. (2010) had participants learn to categorize images of bird species and manipulated whether training included some testing events (i.e., indicating which species new exemplars belong to) or only study opportunities (i.e., receiving the exemplars with the species name present). The testing group later outperformed the study-only group when categorizing new exemplars. Therefore, at least for simple learning tasks, there is good evidence that retrieval practice can be used during training to enhance transfer outcomes.

There is also some evidence that the benefits of retrieval practice extend to transfer tasks in authentic educational settings (Larsen et al., 2013; McDaniel et al., 2007; for review, see Agarwal et al., 2012). For example, McDaniel et al. (2007) introduced weekly quizzes in online science courses and compared this intervention to a control condition that only involved extra reading. Later, for final exam questions that tapped application of course concepts, students who had received weekly quizzes outperformed those in the study-only control group. Most

convincingly, Larsen et al. (2013) found that medical students who had undergone retrieval practice while learning clinical topics (relative to a study-only control condition) were better able to transfer their knowledge to encounters with standardized patients six months later. Unresolved issues about task complexity notwithstanding[3], these studies suggest that retrieval practice can enhance subsequent transfer under some conditions.

The learning gains caused by retrieval practice are often explained with reference to the difficult (e.g., Bjork & Bjork, 2011), effortful (e.g., Carpenter & DeLosh, 2006; Pyc & Rawson, 2009), or elaborative (e.g., Carpenter, 2009; 2011; Rawson et al., 2015) nature of retrieval attempts. More specifically, having to retrieve studied information during training is thought to activate other related information in memory, thereby enhancing retention of the target information because it can be accessed via more retrieval routes (for reviews, see Karpicke et al., 2014; Kornell & Vaughn, 2016; c.f. Lehman & Karpicke, 2016). Indeed, forms of retrieval practice that provide weaker retrieval cues (e.g., free recall) are known to enhance retention more so than forms that provide stronger retrieval cues (e.g., recognition), because weaker retrieval cues presumably demand a more effortful or elaborate search through memory to bring to mind the target information (e.g., Butler & Roediger, 2007; Carpenter & DeLosh, 2006; Carpenter & Yeung, 2017; McDaniel et al., 2007; for meta-analytic review, see Rowland, 2014),

Contextual Variability

Lastly, it is widely acknowledged that varied conditions of learning—what I will broadly refer to as *contextual variability*—can enhance subsequent transfer, but the key variables at play are poorly understood. The main idea dates at least as far back as William James (1899): "[T]he same thing recurring on different days, in different contexts, read, recited on, referred to again and again . . . gets well wrought into the mental structure. (p. 80)". It also seems intuitive that encountering multiple different contexts during training will promote transfer to new contexts. What remains unclear is when this logic holds true.

Even for experiments investigating the retention of simple learning materials (e.g., words and word pairs), some studies show a clear benefit of varied contexts across encoding events

[3] Some authors are less convinced about the merits of retrieval practice for more complex learning tasks, arguing its benefits diminish as task complexity increases (e.g., van Gog & Sweller, 2015). This idea remains contentious, however, because it is not clear how to quantify task complexity (e.g., Karpicke & Aue, 2015; Rawson, 2015).

(e.g., Gartman & Johnson, 1972; Glenberg, 1979; Cuddy & Jacoby, 1982; Melton, 1970) whereas many other studies do not (e.g., Greene & Stillwell, 1995; Postman & Knecht, 1983; Verkoeijen et al., 2004; Slamecka & Barlow, 1979). For example, Slamecka and Barlow (1979) had participants encode homographs that were shown twice, in either identical contexts (tusk–horn, tusk–horn), similar contexts (tusk–horn, antler–horn), or dissimilar contexts (e.g., tusk–horn, whistle–horn). These authors found that the identical contexts led to better performance than did dissimilar or similar contexts, contrary to the idea that contextual variability promotes learning. Contextual variability thus seems to enhance simple forms of learning under some circumstances, but the key variables at play are not well understood (for a more detailed discussion of the mixed results, see Delaney et al., 2010).

The effect of contextual variation during training also remains unclear for more complex problem-solving tasks. This issue has typically been studied in the analogical transfer literature by considering the surface resemblance between two or more training examples, as contextual variability can be conceptualized based on how much examples differ in their particular surface features (see Day & Goldstone, 2012, p. 159). Yet studies reveal mixed findings on this front. For example, Halpern et al. (1990) had participants read passages on a science topic that were either superficially similar or dissimilar in their surface features and found the dissimilar examples led to better transfer (also see Rittle-Johnson & Star, 2009; Holyoak & Koh, 1983). However, as noted by Day and Goldstone (2012), "the less similar two situations are overall, the less likely it becomes that corresponding entities will share overt surface similarities, and thus the process of mapping itself becomes both more cognitively demanding and prone to error. (p. 159)" Accordingly, Kotovsky and Gentner (1996) found that learners had difficulty aligning the key elements of perceptually different examples. Examples with similar surface features can also be harmful because irrelevant details may be retained in the learner's representation of the principle at the expense of more general structure (e.g., Vokey & Brooks, 1992; also see Medin & Ross, 1989). It thus remains unclear when and how contextual variability during learning enhances subsequent transfer.

One explanation for these mixed findings is that, at least for problem-solving tasks, contextual variability is almost always studied through the lens of encoding processes (e.g., comparing dissimilar examples presented simultaneously) and seldom through the lens of retrieval processes (c.f. Butler et al., 2017). For instance, based on the retrieval practice literature

the changing of contexts or surface features across training problems might provide weaker retrieval cues with which to access prior examples that are stored in memory (e.g., Cuddy & Jacoby, 1982; Jacoby, 1978), thereby increasing the effort or depth of retrieval required to solve the problem. These increased retrieval demands introduced through contextual variability might then make one's training experiences more memorable, resulting in better learning and transfer. I tested this hypothesis experimentally in Chapter 2.

Correlational Approaches

In contrast to the experimental approach, relatively little work has sought to identify individual differences in students that correlate with subsequent transfer outcomes. Nonetheless, I summarize some key findings below, mainly focusing on a popular self-report inventory that is meant to predict a student's tendency to learn so that they can later generalize their understanding of key principles to new contexts.

Basic Concept-Learning Tasks

Some success has come from basic laboratory tasks that classify students based on two distinct types of learning. So-called "exemplar learners" are said to focus on the concrete specifics of examples, whereas "abstraction learners" are said to focus on underlying rules or principles conserved across examples (Frey et al., 2017; McDaniel et al., 2014; McDaniel et al., 2018). For example, McDaniel et al. (2014) had participants complete a simple function-learning task where they first received several training exemplars (i.e., *x* values and associated *y* values from a bi-linear "V" function, which participants never saw directly) and then were asked to predict the *y* values that correspond to new *x* values. There was a clear dissociation such that some participants classified as abstraction learners reliably gave *y* values matching the bi-linear function, whereas other participants classified as exemplar learners did not, thus suggesting a difference in how participants processed the training examples. These learner profiles also predicted performance on two other categorization tasks (McDaniel et al., 2014, Exp 1c, 2) and performance on application questions, but not memorization questions, in university chemistry courses (McDaniel et al., 2018). Taken together, these findings support the stability and generalizability of these learner profiles to authentic educational settings.

Student Learning Inventories and the SPQ

Far more attention has been directed toward student learning inventories within the education literature. Built upon theory from the seminal Student Approaches to Learning framework (e.g., Biggs, 1987; Entwistle et al, 1979; Marton & Säljö, 1976; Pask & Scott, 1972), these inventories all assume that the specific learning approaches adopted by students can be measured and used to predict the types of learning outcomes that will emerge as a result from their studies. Most relevant for present purposes, the discussion invariably revolves around the process of *deep learning* as a way to promote transfer. Definitions of deep learning vary slightly by author (e.g., Barnett & Ceci, 2002; Chernobilsky et al., 2004; Entwhistle & McCune, 2004; Sandberg & Barnard, 1997), but apt examples include the processing of content "at a high level of generality, such as main ideas, themes, and principles" (Biggs, 1993, p. 7), and "understanding core concepts [to see] relationships among them or figuring out how to apply information in new ways" (Laird et al., 2014, p. 405). Such descriptions suggest these inventories were designed to predict a student's tendency to understand core principles and transfer that knowledge to new situations.

The Study Process Questionnaire (SPQ; Biggs et al., 2001) is arguably the most popular of these inventories for measuring deep learning. It simply prompts a student with Likert items about their study attitudes and habits within a course or curriculum, in turn producing a "deep approach" score and "surface approach" score. Thus, the main assumption behind the SPQ is that its deep approach measure reflects a student's tendency to extract general principles and transfer them to solve new problems, at least within a given educational setting. It is easy to see why such a tool for quantifying deep learning in practice would be of interest to educators and educational developers.

Despite the SPQ's intuitive appeal and widespread use in education (for review, see Baeten et al., 2010), its predictive validity remains unclear. Two major criticisms stand out: first, there is much ambiguity in what specific learning outcomes serve as evidence of deep learning; and second, there is a scarce empirical evidence to show that such outcomes positively correlate with SPQ deep approach scores (Dinsmore & Alexander, 2012; Howie & Bagnall, 2013). Amplifying these criticisms, a large meta-analysis recently found that deep approach scores correlated only very weakly with overall academic achievement in university students (r^+ = .14,

p = .03), accounting for less than 3% of variance (Richardson et al., 2012). These modest findings cast doubt on the utility of the SPQ for its intended purpose in practice.

However, such weak associations might be an artefact of SPQ studies largely relying on aggregate outcome measures (e.g., final exam performance, cumulative GPA) that are highly heterogenous—and hence may not precisely reflect whether students extracted general principles from the courses under investigation (see Watkins, 2001, as cited in Choy et al., 2012). Therefore, what remains ambiguous is whether SPQ deep approach scores positively correlate with outcome measures that more precisely index how well students can transfer their understanding of key principles. I address this issue using a correlational approach in Chapters 3 and 4, examining SPQ scores in relation to student performance on transfer tasks in both the laboratory and the classroom.

Of secondary interest, another measurement issue is that SPQ scores are thought to be highly specific to the learning environment for which the tool was administered, but available data do not completely support this view. The claim is that SPQ scores obtained from students are sensitive to the specifics of a given learning environment, including variables like the content being learned, the instructor's style of assessment, prior knowledge, and the student's perceptions of learning objectives (e.g., Biggs, 1993; Baeten et al. 2010; Struyven et al. 2006; Wilson & Fowler, 2005). It is therefore considered inappropriate to generalize a student's SPQ scores obtained in one educational context to how they will approach studying in other educational contexts (e.g., Biggs et al., 2001). Nonetheless, some studies report substantial overlap between learning approaches and measures of personality (Duff et al., 2004; Chamorro-Premuzic & Furnham, 2009; Chamorro-Premuzic et al., 2007; Zhang, 2003) or stable individual differences in information processing (Bouckenooghe et al., 2016), implying deep approaches may be trait-like. Informal inspection of SPQ items also suggests that some of the targeted constructs might reflect general traits that persist across different learning contexts (e.g., "I feel that virtually any topic can be highly interesting once I get into it, and "I spend a lot of my free time finding out more about interesting topics which have been discussed in different classes"). I addressed this issue of the SPQ's context specificity in Chapter 3.

Chapter 2: A synergistic effect of contextual variability and spacing on subsequent transfer

The present line of research was inspired by two literatures on analogical transfer and retrieval-based learning. With respect to analogical transfer, the learning benefits of contextual variability are usually investigated such that an experimental group is shown multiple examples of the same principle simultaneously and asked (or guided) to compare them, whereas a control group is shown each example in isolation without any comparison instructions. Much work has shown that this comparison process leads to better subsequent transfer than the non-comparison control condition. This effect is usually explained by invoking the *variability-induced encoding* hypothesis, which suggests that comparison facilitates encoding of the higher-order structure shared by the examples.

In contrast, with respect to retrieval-based learning, the impact of contextual variability is conceptualized based on increased retrieval demands across successive test events, as retrieval cues are presumably weaker when tested in new contexts. These increased retrieval demands via contextual variability are in turn thought to promote subsequent retention and transfer—an idea I refer to as the *variability-induced retrieval* hypothesis. However, support for this hypothesis stems mostly from simple memory experiments involving word lists and verbatim recall. In this chapter I therefore tested the latter retrieval-based learning hypothesis using a task more reminiscent of analogical transfer.

Note how these two hypotheses have separate but equally important implications for educational practice. The variability-induced encoding hypothesis suggests that, at the time of initial learning, there is value in placing multiple examples with varying surface features side-by-side to facilitate encoding of their deep structure. However, we must also remember that learning is not a one-time event; it is often distributed over longer periods of time, like throughout a course or curriculum, such that the content is periodically retrieved and built upon. In that vein, the variability-induced retrieval hypothesis suggests that successive formative assessments ought to be varied in their contexts of presentation, imposing greater retrieval demands that promote subsequent retention and transfer.

A synergistic effect of contextual variability and spacing on subsequent transfer

Andrew B. LoGiudice[1], Geoffrey R. Norman[2], Sandra Monteiro[2],

Kulamakan M. Kulasegaram[3], & Scott Watter[1]

[1]McMaster University, Department of Psychology, Neuroscience & Behaviour

[2]McMaster University, Department of Health Research Methods, Evidence and Impact

[3]University of Toronto, Department of Family & Community Medicine

Correspondence addressed to Andrew B. LoGiudice, Department of Psychology, Neuroscience & Behaviour, McMaster University, 1280 Main Street West, Hamilton, ON L8S-4K1. Email: logiudab@mcmaster.ca

Abstract

Theorists often assume that varied conditions of training give rise to more robust learning. What arguably remains unclear is *when* and *how* different types of contextual variability improve learning depending on the task at hand. Here we explored this issue with respect to analogical transfer—i.e., the process by which we learn abstract principles through examples. Prior work suggests that variability in surface features across examples can improve learning by helping the learner encode conserved deep structural characteristics, which we refer to as variability-induced encoding. In contrast, here we hypothesized that variability in surface features across a set of to-be-solved training problems requires more effortful or 'deep' retrieval of relevant prior training events, thereby enhancing learning and subsequent transfer. We refer to this as the variability-induced retrieval hypothesis. Consistent with this latter hypothesis, we found a positive effect of contextual variability on subsequent transfer when we encouraged retrieval by presenting to-be-solved training problems spaced in time, but not when we discouraged retrieval by giving participants access to earlier problems and their solutions throughout training. These findings suggest contextual variability can sometimes improve transfer outcomes not by enhancing encoding of deep structure, but by encouraging more effortful retrieval during training. We discuss these findings in relation to analogical transfer and retrieval-based learning.

Keywords: Analogical Transfer; Contextual Variability; Desirable Difficulties; Retrieval Practice; Retrieval-based Learning

Introduction

Theorists have long argued that contextual variability during training promotes robust and flexible learning. This simple idea harks back to William James (1899, p. 80), "[T]he same thing recurring on different days, in different contexts, read, recited on, referred to again and again . . . gets well wrought into the mental structure", and yet remains a cornerstone in discussions about learning and memory (e.g., Bransford & Schwartz, 1999; Karpicke et al., 2014; Smith & Handy, 2014; Posner & Keele, 1968). It also seems intuitive that performing a task in many different contexts will lead to more robust learning than performing the task repeatedly in the same context. For instance, when teaching Pythagoras' theorem, it seems wise to have students apply their knowledge to various problems that differ in whether the hypotenuse or another side length must be deduced, whether a right-angled triangle or isosceles triangle is shown, and so on. But what arguably remains unclear is *when* or *how* contextual variability promotes learning, and how this depends on the specific characteristics of a given learning task.

Here we explored the effect of contextual variability on analogical transfer—i.e., the ability to extract general principles from examples and solve analogous problems (for reviews, see Day & Goldstone, 2012; Hager & Hodkinson, 2009; Reeves & Weisberg, 1994)—based on theories that posit a key role for effortful, elaborative, or 'deep' retrieval of information from memory (e.g., Karpicke, 2012; Pyc & Rawson, 2009; Rawson & Dunlosky, 2011; Roediger & Butler, 2011). Specifically, we hypothesized that variability in surface features across practice problems that are spaced in time might increase retrieval demands, thereby enhancing learning and subsequent transfer. In what follows, we outline a rationale for this hypothesis and report two experiments designed to test it.

Context Specificity

Transfer researchers usually describe 'context' in terms of surface features; that is, the superficial characteristics of examples that illustrate a general principle in some concrete way (Day & Goldstone, 2012; Reeves & Weisberg, 1994). For example, a tutor might teach Newton's second law by presenting several examples of vector diagrams that depict spacecrafts generating thrust. In this case surface features might refer to the specific objects depicted, the magnitudes or angles of the vectors, the semantic domain (e.g., physics, spaceflight), or any other incidental characteristics of the examples.

These surface features may seem arbitrary, but they are vital to studies of transfer, because a learner's likelihood of solving a new problem depends heavily on its surface resemblance to previously seen examples (Gentner & Markman, 1997; Gick & Holyoak, 1987, 1983; Holyoak & Koh, 1987; Keane, 1987; Ross, 1987). Surface features of examples are thus said to be inextricably linked with a learner's understanding of the underlying principle, making it hard for them to infer how the principle generalizes to unfamiliar contexts. Therein lies the dilemma: how can we help students overcome this context specificity by structuring learning in a way that is most conducive to transfer?

Variability-Induced Encoding

One potent way to improve transfer outcomes is to present multiple examples of the same principle that differ in their surface features, in part because this helps the learner to encode deep structural relations conserved across examples (Day et al., 2010; Gick & Holyoak, 1983; Goldstone & Sakamoto, 2003; Markman & Gentner, 2000; Rittle-Johnson & Star, 2009). We will refer to this idea as the *variability-induced encoding hypothesis*.

Though fruitful, this emphasis on enhanced encoding has overshadowed potential retrieval-based mechanisms. Indeed, given the recent surge of studies showing that recall and transfer outcomes can be improved via retrieval demands during training (for reviews, see Karpicke et al., 2014; Roediger & Butler, 2011; for meta-analysis, see Rowland, 2014), it may also be useful to investigate analogical transfer by examining the interplay between contextual variability and retrieval processes.

Variability-Induced Retrieval

In particular, it seems plausible that contextual variability during training provides weaker retrieval cues with which to access relevant past events, enhancing subsequent retention (e.g., Jacoby, 1978). To illustrate the same idea, Cuddy and Jacoby (1982, Exp. 3) had participants "solve" a series of paired associates (e.g., Lawyer C--rt) by naming aloud each righthand target word. Critically, some pairs were repeated: either the same letters of the target (Lawyer C--rt) or different letters of the target (Lawyer -our-) were deleted. The latter trials, which resemble a form of contextual variability, were thought to "reduce the effectiveness of the second presentation of the problem as a cue for retrieval of the solution constructed on its first presentation", which would require "further processing at the time of the second presentation of

a problem to arrive at a solution, and consequently, enhance subsequent retention" (p. 458). Final cued recall performance aligned with this reasoning. Similarly, increasing retrieval demands through other means (e.g., lag, impoverished retrieval cues) is generally thought to improve retention (e.g., Carpenter, 2009; Pyc & Rawson, 2009; Whitten & Bjork, 1977). Together these findings suggest contextual variability during training can enhance learning through increased retrieval demands—what we will call the *variability-induced retrieval hypothesis*.

Variability-Induced Retrieval and Transfer?

We wondered whether this retrieval-based logic scales up to more complex transfer tasks. Imagine a student learning a physics principle by solving two practice problems spaced a few minutes apart. Let us assume they successfully solve the first problem and the experience is memorable. However, if the second problem has nearly identical surface features, the student might trivially solve it by accessing their construction of the first problem's solution from memory. In this case, because the second problem's solution was guided by a similar prior experience and unfolded so fluently, one might say minimal retrieval demands were imposed. In contrast, if the second problem has very different surface features than the first problem, the learner might instead be forced to effortfully or 'deeply' recollect what they remember about the first problem and map that information onto the unfamiliar surface features of the second problem. In this case, according to the variability-induced retrieval hypothesis, the retrieval demands introduced through contextual variability ought to make the student's training experiences more memorable, thereby enhancing subsequent transfer of the principle.

An experimental design used by Kulasegaram et al. (2017) offers a way to test this retrieval-based hypothesis. These authors had participants learn three physiological principles, and contextual variability was manipulated by presenting to-be-solved training problems for each principle in either multiple organ system contexts (e.g., respiratory, digestive, cardiovascular) or a single organ system context (e.g., only respiratory). The key finding was that the multiple-context group outperformed the single-context group when later asked to solve analogous problems embedded in novel organ system contexts (e.g., urinary). The authors explained this finding by invoking the variability-induced encoding hypothesis, implying the benefit of contextual variability was driven by enhanced encoding of deep structural relations conserved across training problems. However, the variability-induced retrieval hypothesis suggests the

benefit of contextual variability may have instead stemmed from increased retrieval demands. We therefore modified this design to test the variability-induced retrieval hypothesis.

Present Studies

In Experiment 1 we sought to replicate the finding of enhanced transfer when contextual variability is imposed across a set of to-be-solved training problems that are presented 'spaced' in time. The spaced nature of the task was also exaggerated by inserting filler tasks between successive training problems. Variability-induced encoding might be operative in that dissimilar surface features encourage learners to focus on the deep structural relations conserved across problems. But, because this spaced format encourages learners to recall prior problems and their solutions as a means to solve future training problems, variability-induced retrieval might also be operative in that dissimilar surface features across training problems impose greater retrieval demands. Thus, we assumed any benefit of contextual variability that emerges under these conditions can be attributed to either mechanism.

In Experiment 2 we then explored an 'adjacent' training format where, akin to an open-book test, previously solved training problems and their correct solutions remained visible to participants while they solved all remaining training problems for the same principle. Variability-induced encoding would still presumably be operative (or even more pronounced[4]) under these circumstances for the same reason it would be operative in Experiment 1. But, relative to the spaced format of Experiment 1, we reasoned that this adjacent format would reduce or eliminate the need to engage in effortful retrieval of prior training events because prior training problems and their solutions remained visible throughout training. Thus, we assumed any benefit of contextual variability that emerges under these conditions where variability-induced retrieval is stifled would selectively reflect the impact of variability-induced encoding.

Most importantly, by comparing across experiments, we assumed that any additional positive effect of contextual variability obtained in Experiment 1 (when retrieval is encouraged

[4] Placing multiple examples of the same principle side-by-side is generally thought to promote identification of deep structural relations (see Day & Goldstone, 2012), and so a strong interpretation of the variability-induced encoding hypothesis predicts that the positive effect of contextual variability should actually be *more* pronounced in the massed format than the spaced format.

through the spaced format) relative to Experiment 2 (when retrieval is discouraged through the adjacent format) would serve as evidence for the variability-induced retrieval hypothesis.

Experiment 1

Method

Design

The design consisted of a between-participant contrast between two groups that differed only in the problems received during training. The *single-context* group (n = 30) received training problems such that each principle was instantiated in a single organ system context, whereas the *multiple-context* group (n = 30) received training problems such that each principle was instantiated in three different organ system contexts. Both groups then received a series of analogous problems to solve in the test phase as a final measure of transfer.

Participants

A total of 60 undergraduate students (Mean Age = 19.6, 72% Female) at McMaster University participated for course credit. The study was approved by the McMaster Research Ethics Board (Protocol #1955), and all participants gave informed consent.

Materials

Training and test materials were adopted based on those by Kulasegaram et al. (2017). Briefly, these materials centered around three physiological principles: Laplace's law describes the relation between radius, pressure, and tension in a cylindrical vessel; turbulent flow, fictitiously labeled Goethe's law for consistency, describes the relation between flow, velocity, viscosity, radius, and roughness in a cylindrical vessel; and Starling's law describes the relation between stroke volume and ventricular stretching in the heart. For ease of exposition, we describe the nature of these training and test materials throughout the Procedure section below.

Procedure

Participants were seated at individual computer booths in a quiet laboratory setting. Learning materials were shown on a computer monitor and all written responses were typed using a standard keyboard. The entire experiment was programmed using PsychoPy2 software (Peirce, 2007).

During the training phase, participants were asked to learn the three principles for an upcoming test that would require them to solve new problems. Principles were learned in a blocked fashion (i.e., all training materials for a principle were received before learning the next principle), and training was self-paced with the constraint that participants could not go back to any training material once they chose to proceed past it. For each principle, the participant began by reading a passage that described the principle[5]. A key was then pressed to proceed to the first training problem. A sample passage and training problem are shown in Appendix A. Once the participant submitted their solution attempt, it was replaced by a normative description of the solution, which they could examine for as long as they wished. The participant then pressed a key to remove all this information from the display and proceed to the next training problem. This process repeated for three training problems. Then training for the next principle began. Order of principles and training problems was randomized for each participant. To reinforce the spaced nature of the task, participants also performed a 45 s filler task before each training problem where they received a unique category label (e.g., fruits) and listed category members.

Contextual variability was manipulated during training by providing different sets of training problems to the two groups. The single-context group received three Laplace training problems in the context of the digestive system and three Goethe training problems in the context of the respiratory system[6]. In contrast, the multiple-context group received one Laplace training problem in each of the digestive, cardiovascular, and lymphatic systems, and one Goethe training problem in each of the respiratory, cardiovascular, and nervous systems. Contextual variability was thus operationally defined based on the number of organ systems in which training problems for a given principle were instantiated.

During the subsequent test phase, participants then immediately tried to solve 13 test problems shown in sequence. Order of test problems was randomized for each participant. The participant's task for each test problem was to decide which of the three principles best applied and to write a brief solution attempt that justified their decision. They then pressed a key to

[5] Laplace's law was described in the context of the digestive system, Goethe's law in the context of the respiratory system, and Starling's law in the context of the cardiovascular system.

[6] Both groups received the same three training problems for Starling's law. This principle is not relevant to the manipulation of contextual variability; it was included only to keep the design comparable to that of Kulasegaram et al. (2017).

submit their solution attempt and proceed to the next problem. No feedback was provided after submitting each solution attempt.

Scoring

Similar to Kulasegaram et al. (2017), we differentiated between two measures of transfer in the test phase. Seven test problems (Laplace = 2, Goethe = 2, Starling = 3) were designed to probe understanding of the principle in a familiar context—i.e., an organ system the participant previously encountered for the same principle, which was always the same organ system that the single-context group had received for all training problems of that principle. We labelled these near transfer test problems because we assumed the match in context would facilitate retrieval of prior training events for the appropriate principle, especially for the single-context group who received all relevant training problems in that context. The remaining six test problems (Laplace = 3, Goethe = 3) were instantiated in novel organ system contexts that neither group had encountered; we labelled these far transfer test problems. We disentangled these two measures of transfer because we assumed the far transfer test problems would provide a more sensitive and purer measure of learning than near transfer test problems.

Solution attempts for training and test problems were all scored from 0–3 with a higher score reflecting greater understanding of the correct principle (Kulasegaram et al., 2017). The scoring scheme was as follows: 0 = Does not capture correct principle; 1 = Captures correct principle but with obviously flawed understanding; 2 = Captures correct principle but with key ideas absent or misconstrued; 3 = Captures correct principle with all key ideas present and described accurately.

To ensure reliability in scoring, training problem solution attempts from a subset of 25 participants were first scored by both the lead author and an independent rater naive to the purpose of the study. Inter-rater reliability was high (ICC = .85), so all training problem solution attempts reported hereafter were scored by the independent rater. Similarly, solution attempts for test problems from a subset of 55 participants were scored by both the lead author and an independent rater while blinded to group membership. Inter-rater reliability was also high for both near transfer test problems (ICC = .90) and far transfer test problems (ICC = .89), so all test problem solution attempts reported hereafter were scored by the lead author in a blinded fashion.

Results

Exclusions

Data from two participants were excluded because of software failure, and data from one participant were excluded because more than half of their test responses were left blank, suggesting noncompliance. This left a total of 57 participants in the following analyses (single-context group, $n = 28$; multiple-context group, $n = 29$).

Training Phase

Mean training problem scores were submitted to an independent samples *t*-test with contextual variability as the independent variable. There was no significant difference between groups, $t_{(55)} = 1.49$, $p = .143$, though performance was numerically greater for the single-context group (M = 1.84, SEM = .12) than the multiple-context group (M = 1.60, SEM = .11).

Test Phase

Mean test problem scores were submitted to a 2x2 mixed-design ANOVA with contextual variability (single context, multiple contexts) as a between-participant factor and problem type (near transfer, far transfer) as a within-participant factor. These data are shown in Figure 1. There was a significant main effect of contextual variability, $F_{1,55} = 10.9$, $p = .002$, $\eta_p^2 = .166$, such that the multiple-context group outperformed the single-context group. This finding constitutes a replication of the key result observed by Kulasegaram et al. (2017). There was also a significant main effect of problem type $F_{1,55} = 58.7$, $p < .001$, $\eta_p^2 = .516$, such that performance was greater overall for near transfer test problems than far transfer test problems. The interaction was not statistically significant, $F < 1$.

< Figure 1 here >

Discussion

The training phase data were somewhat surprising because we expected there to be a larger effect of contextual variability. That is, if consistency in context across training problems facilitates retrieval of prior training problems, one might expect the single-context group to produce higher quality solution attempts for the training problems. Yet there was only a small numerical trend in this direction. We elaborate on this finding after Experiment 2.

The near transfer data were also surprising because they differ from the findings of Kulasegaram et al. (2017). Whereas their data show that the single-context group outperformed the multiple context group on near transfer test problems, our data showed the opposite pattern. Means for the single-context group are comparable between our experiment (M = 1.26) and theirs (M = 1.28), indicating that the discrepancy is driven by the multiple-context group performing better in our experiment (M = 1.78) than in their experiment (M = 1.07). We speculate that this discrepancy may have arisen because we exaggerated the spaced format by inserting filler tasks between training problems. That is, if our exaggerated spaced format imposed greater retrieval demands, and these retrieval demands promote learning, this would explain why we observed a more pronounced positive effect of contextual variability for near transfer problems.

Experiment 2

The purpose of Experiment 2 was to explore whether this positive effect of contextual variability on subsequent transfer would be moderated by reduced retrieval demands in the adjacent format. To reiterate, we were most interested in a potential interaction across experiments such that the benefit of contextual variability on subsequent transfer would be more pronounced when paired with the spaced format in Experiment 1 than when paired with the adjacent format in Experiment 2. Such an interaction would suggest that the benefit of contextual variability observed in Experiment 1 was caused in part by greater retrieval demands. Therefore, in addition to analyzing the data from Experiment 2 in isolation, we also planned to conduct an analysis with the combined test phase data from Experiments 1 and 2. Specifically, we planned to submit the combined data to separate 2x2 ANOVAs for near transfer performance and far transfer performance, with contextual variability (single context, multiple context) as one between-participant factor and training format (Experiment 1: spaced, Experiment 2: adjacent) as a second between-participant factor.

Given the numerical trend toward better training problem performance for the single-context group than the multiple-context group in Experiment 1, we also planned to submit the training phase data from both experiments to a 2x2 ANOVA with contextual variability (single context, multiple context) as one between-participant factor and training format (Experiment 1: spaced, Experiment 2: adjacent) as a second between-participant factor.

Method

Participants

A total of 60 undergraduate students (Mean Age = 19.0, 72% Female) at McMaster University participated for course credit. The study was approved by the McMaster Research Ethics Board (Protocol #1955), and all participants gave informed consent. Thirty participants were randomly assigned to the single-context group and 30 to the multiple-context group.

Procedure

The procedure was nearly identical to that of Experiment 1 except that training problems and their associated solution feedback remained visible while the participants solved all ensuing training problems for the same principle. We also removed the 45-second filler task before each training problem to further exaggerate the massed nature of the adjacent training format.

Results

Exclusions

Data from one participant were excluded because of software failure, and data from one participant were excluded because more than half of their test responses were left blank, suggesting noncompliance. One extra participant was also accidentally recruited in the single-context group. This left a total of 59 participants in the following analyses (single-context group, $n = 31$; multiple-context group, $n = 28$).

Training Phase

Mean training problem scores were submitted to an independent samples *t*-test with contextual variability as the independent variable. There was no significant difference between groups, $t_{(57)} = 1.51$, $p = .138$, though performance was again numerically greater for the single-context group (M = 1.95, SEM = .12) than the multiple-context group (M = 1.71, SEM = .10).

Test Phase

Mean test problem scores were submitted to a 2x2 mixed-design ANOVA with contextual variability (single context, multiple contexts) as a between-participant factor and problem type (near transfer, far transfer) as a within-participant factor. These data are shown in Figure 2. In contrast to Experiment 1, there was no significant main effect of contextual variability, $F_{1,57} = 2.31$, $p = .134$, $\eta_p^2 = .039$. Like Experiment 1, however, there was a significant

main effect of problem type $F_{1,57} = 102.8$, $p < .001$, $\eta_p^2 = .643$, such that performance was greater overall for near transfer test problems than far transfer test problems. The interaction was not statistically significant, $F < 1$.

< Figure 2 here >

Combined Analysis of Experiments 1 and 2

Training Phase

Training problem solution attempts from both experiments were submitted to a 2x2 ANOVA with contextual variability (single-context, multiple-context) as one between-participant factor, and training format (Experiment 1: spaced, Experiment 2: adjacent) as a second between-participant factor. These data are shown in Figure 4. This combined analysis revealed a significant main effect of contextual variability such that, overall, better solutions were generated by the single-context groups than the multiple-context groups, $F_{(1,112)} = 4.47$, $p = .037$, $\eta_p^2 = .038$. The main effect of the training format was not significant, nor was the interaction, Fs ≤ 1.

< Figure 3 here >

Test Phase

The data from Experiments 1 and 2 were also submitted to separate 2x2 ANOVAs, one for near transfer performance and one for far transfer performance, with contextual variability (single context, multiple context) as one between-participant factor and training format (Experiment 1: spaced, Experiment 2: adjacent) as a second between-participant factor. These data are shown in Figure 4.

< Figure 4 here >

For near transfer performance, the combined analysis revealed a significant interaction between contextual variability and training format, $F_{(1,112)} = 10.9$, $p = .001$. The main effects of training format, $F_{(1,112)} < 1$, and contextual variability, $F_{(1,112)} = 1.39$, were not statistically significant. Inspection of the left panel in Figure 4 suggests that the interaction was driven by a substantial advantage of the multiple context group over the single-context group when training problems were spaced, but also by a slight advantage of the single-context group over the multiple context group when training problems were solved adjacent to one another.

For far transfer performance, the combined analysis revealed a main effect of training format such the sequential format (M = .91) produced greater performance overall than the adjacent format (M = .72), $F_{(1,112)} = 3.94$, $p = .05$, $\eta_p^2 = .034$. However, this main effect was qualified by a significant interaction between contextual variability and training format, $F_{(1,112)} = 6.69$, $p = .011$. Inspection of the right panel in Figure 4 shows this interaction was driven by a large advantage of the multiple context group over the single-context group when training problems were spaced in Experiment 1. The main effect of contextual variability was not statistically significant, $F_{(1,112)} = 2.00$, $p = .160$.

Expecting that the retrieval demands imposed by the spaced format might equate to a longer time needed to solve training problems relative to the adjacent format, we also explored whether the time taken to solve training problems differed as a function of training format[7]. These data are shown in Figure 5. Contrary to our expectations, the 2x2 ANOVA revealed a significant main effect of training format, $F_{(1,112)} = 6.99$, $p = .009$, $\eta_p^2 = .059$, such that participants undergoing the spaced format (M = 2.50 min) actually took less time to generate solution attempts per training problem than participants undergoing the adjacent format (M = 3.02 min). Neither the main effect of contextual variability nor the interaction was statistically significant, Fs < 1.

< Figure 5 here >

General Discussion

Contextual variability during training improved transfer outcomes when the training problems were spaced in time (i.e., *without* direct access to prior training problems or their solutions), but not when training problems were solved adjacent to one another (i.e., *with* direct access to prior training problems and their solutions). This finding therefore constitutes a synergistic effect of contextual variability and spacing on subsequent transfer.

No Evidence of Variability-Induced Encoding

The present data are inconsistent with the variability-induced encoding hypothesis. This hypothesis posits that variability in surface features across a set of training problems can help the

[7] Time taken to complete filler tasks for the spaced format was excluded in this analysis

learner encode conserved deep structural relations, thus predicting contextual variability ought to be beneficial when paired with either the spaced or adjacent format. If anything, because placing examples side-by-side is generally thought to promote comparison and identification of deep structural relations (see Day & Goldstone, 2012), a strong interpretation of this hypothesis predicts that the positive effect of contextual variability ought to be more pronounced for the adjacent format than the spaced format. Contrary to these predictions, contextual variability did not enhance subsequent transfer at all when paired with the adjacent format[8].

One explanation for this finding is that participants were never given explicit instructions to identify similarities between training problems. For example, Catrambone and Holyoak (1989) concluded that the specific instructions participants received while studying multiple examples had a large impact on subsequent transfer, arguing that explicit, guided comparison of examples was required for robust transfer because "novices in a domain tend to focus on those features of examples with which they are most familiar and to miss the underlying concepts that the examples are supposed to demonstrate" (p. 1154). Similarly, much work has shown that transfer outcomes are enhanced when learners are explicitly told to compare multiple examples relative to when they instead receive each example in isolation without such comparison instructions (e.g., Catrambone & Holyoak, 1989; Christie & Gentner, 2010; for meta-analysis, see Alfieri et al., 2013; cf. Quilici & Mayer, 2002; Ross & Kennedy, 1990). Thus, we may not have detected a positive effect of contextual variability within the adjacent format because we did not *explicitly* guide participants to focus on the deep structural relations conserved across training problems.

Evidence of Variability-Induced Retrieval

The present data are more consistent with the variability-induced retrieval hypothesis. This hypothesis posits that variable surface features across a set of to-be-solved training problems *that are spaced in time* ought to increase retrieval demands. That is, if a set of training problems have different surface features and are spaced in time, the learner might be forced to more effortfully or deeply recollect what they remember about prior training problems and map that information onto the unfamiliar surface features of the next training problems, consequently

[8] This hypothesis does not fit the data obtained here, but it is still useful for explaining other findings from different paradigms. We suspect the specific nature of the learning task determines whether variability-induced encoding is likely to be operative (e.g., whether the content is perceptual or conceptual in nature, how many irrelevant surface features are present). However, a full discussion of moderating variables is beyond our scope.

making training more memorable. We assumed these retrieval demands would manifest in the spaced format where participants were likely to solve training problems by retrieving prior training problems or their solutions, but be reduced or eliminated in the adjacent format where retrieval was unnecessary because the learner had direct access all prior problems and their solutions throughout training. This reasoning is supported by the observed interaction between contextual variability and training format.

Based on this interpretation, it was surprising that we did not observe a stronger main effect of training format in the combined analysis. After all, if the spaced format requires retrieval to a greater extent than the adjacent format, and increased retrieval demands result in more memorable training experiences, one might have expected a larger overall advantage of the spaced format over the adjacent format. This unexpected pattern of results might be explained by considering that the consistency in surface features for the single-context group may have negated the need to effortfully recall prior training problems or their solutions. By this logic, the spaced format paired with a single training context may have been functionally similar to the adjacent format—that is, despite spacing, the solutions to training problems might have been fluently constructed due to their highly similar surface features, eliminating the need to effortfully retrieve prior training problems or their solutions.

Another unexpected finding was that participants undergoing the spaced training format took less time to solve training problems than participants undergoing the adjacent training format. We are uncertain why this was the case. Considering that the spaced format presumably required more effortful retrieval of prior training problems and their solutions to solve ensuing training problems, the opposite finding might be expected. A possible explanation is that only the adjacent format allowed participants to review prior problems and their solutions periodically throughout training, perhaps leading them to engage with the materials for a longer duration despite this not equating to more learning. Supporting this idea, a supplementary analysis on the combined data from both experiments showed that time taken to solve training problems did not correlate with near transfer performance (r = -.03, p = .751) nor far transfer performance (r = .02, p = .853). This interpretation remains speculative, however, and underscores the need to test the variability-induced retrieval hypothesis using control conditions (other than the adjacent format used here) where training instances are not necessarily shown to participants simultaneously.

Limitations

A limitation is that our manipulation of contextual variability relied on different sets of training problems that likely differed in surface features other than those pertaining to organ system per se. Hence, one might argue that the positive effect of contextual variability may have arisen because the training problems given to the multiple-context group somehow better illustrated the underlying concepts than those given to the single-context group. However, if this were true, one would expect a robust main effect of contextual variability such that, regardless of training format, the multiple-context group would produce better transfer outcomes than the single-context group. We clearly did not observe this pattern here. In fact, there was a numerical trend in the opposite direction for the adjacent training format. We therefore favour an explanation related to variability-induced retrieval over this alternative interpretation.

Another limitation is that we only assessed transfer immediately after training. Immediate criterial tasks are standard in the analogical transfer literature, but the retrieval-based learning literature suggests the mnemonic benefits of retrieval are often more pronounced when final tests are administered after several days (for meta-analysis, see Rowland, 2014). Therefore, future work should examine the combined effects of contextual variability and retrieval demands with longer retention intervals separating training and final measures of transfer.

Conclusion

Transfer poses a formidable challenge for researchers and educators alike. Although it may seem self-evident that contextual variability promotes learning and transfer, much remains unknown about the mechanisms at play and how they differ depending on the task being learned. Only when we better understand these mechanisms in relation to more complex learning tasks can we offer educators concrete suggestions about how and when to incorporate contextual variability in their teaching to promote transfer.

Here we made a small step in that direction by drawing links between analogical transfer and retrieval-based learning. While the analogical transfer literature offers complex learning tasks that go beyond rote memorization, the retrieval-based learning literature offers theory to explain the benefits of contextual variability in terms of additional retrieval demands. Combining the strengths of these two literatures, the present work provides preliminary evidence for the variability-induced retrieval hypothesis: contextual variability across a set of to-be-solved

training problems only enhanced subsequent transfer when said training problems were spaced in time. However, we caution readers from drawing strong conclusions based on this single demonstration. Future work must test this hypothesis using other learning tasks and longer retention intervals to further assess its robustness and generalizability.

Acknowledgements

This research was supported in part by a SSHRC Canada Graduate Scholarship (Doctoral) awarded to the first author. We thank Halley Desai and Muqtasid Mansoor for assistance with scoring. We are also grateful to Larry Jacoby for thoughtful input during a visit to McMaster.

References

Alfieri, L., Nokes-Malach, T. J., & Schunn, C. D. (2013). Learning through case comparisons: A meta-analytic review. *Educational Psychologist*, *48*(2), 87–113.

Bransford, J. D., & Schwartz, D. L. (1999). Chapter 3: Rethinking transfer: A simple proposal with multiple implications. *Review of Research in Education*, *24*(1), 61–100.

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(6), 1563–1569

Catrambone, R., & Holyoak, K. J. (1989). Overcoming contextual limitations on problem-solving transfer. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(6), 1147–1156.

Christie, S., & Gentner, D. (2010). Where hypotheses come from: Learning new relations by structural alignment. *Journal of Cognition and Development*, *11*(3), 356–373.

Cuddy, L. J., & Jacoby, L. L. (1982). When forgetting helps memory: An analysis of repetition effects. *Journal of Verbal Learning and Verbal Behavior*, *21*(4), 451–467.

Day, S. B., & Goldstone, R. L. (2012). The import of knowledge export: Connecting findings and theories of transfer of learning. *Educational Psychologist*, *47*(3), 153–176.

Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American Psychologist*, *52*(1), 45–56.1e

Gick, M. L., & Holyoak. K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, *15*, 1–38.

Hager, P., & Hodkinson, P. (2009). Moving beyond the metaphor of transfer of learning. *British Educational Research Journal*, *35*(4), 619–638.

Holyoak, K. J., & Koh, K. (1987). Surface and structural similarity in analogical transfer. *Memory & Cognition*, *15*(4), 332–340.

Jacoby, L. L. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal of Verbal Learning and Verbal Behavior*, *17*(6), 649–667.

James, W. (1899) *On some of life's ideals*. New York: Henry

Karpicke, J. D. (2012). Retrieval-based learning: Active retrieval promotes meaningful learning. *Current Directions in Psychological Science*, *21*(3), 157–163.

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning: An episodic context account. In *Psychology of Learning and Motivation* (Vol. 61, pp. 237–284). Academic Press.

Keane, M. (1987). On retrieving analogues when solving problems. *The Quarterly Journal of Experimental Psychology*, *39*(1), 29–41.

Kulasegaram, K. M., Chaudhary, Z., Woods, N., Dore, K., Neville, A., & Norman, G. (2017). Contexts, concepts and cognition: Principles for the transfer of basic science knowledge. *Medical Education*, *51*(2), 184–195.

Markman, A. B., & Gentner, D. (2000). Structure mapping in the comparison process. *American Journal of Psychology*, *113*(4), 501–538.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*(3), 353–363.

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, *60*(4), 437–447.

Quilici, J. L., & Mayer, R. E. (2002). Teaching students to recognize structural similarities between statistics word problems. *Applied Cognitive Psychology*, *16*, 325–342.

Rawson, K. A., & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? *Journal of Experimental Psychology: General*, *140*(3), 283–302.

Reeves, L., & Weisberg, R. W. (1994). The role of content and abstract information in analogical transfer. *Psychological Bulletin*, *115*(3), 381–400.

Rittle-Johnson, B., & Star, J. R. (2009). Compared with what? The effects of different comparisons on conceptual knowledge and procedural flexibility for equation solving. *Journal of Educational Psychology*, *101*(3), 529–544.

Roediger III, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20–27.

Ross, B. H. (1987). This is like that: The use of earlier problems and the separation of similarity effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(4), 629–639.

Ross, B. H., & Kennedy, P. T. (1990). Generalizing from the use of earlier examples in problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(1), 42–55.

Rowland, C. A. (2014). The effect of testing versus restudy on retention: a meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432–1463.

Smith, S. M., & Handy, J. D. (2014). Effects of varied and constant environmental contexts on acquisition and retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(6), 1582–1593.

Whitten II, W. B., & Bjork, R. A. (1977). Learning from tests: Effects of spacing. *Journal of Verbal Learning and Verbal Behavior*, *16*(4), 465–478.

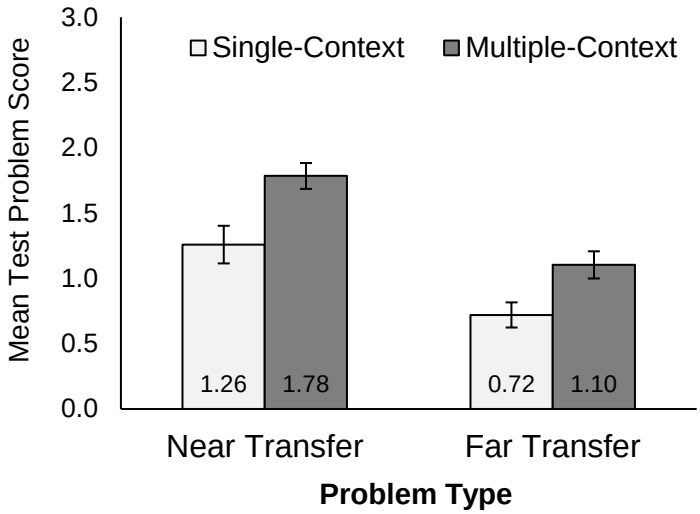


Figure 1. Test problem scores of Experiment 1 as a function of contextual variability (single-context, multiple-context) and test problem type (near transfer, far transfer). These scores reflect the quality of solution attempts for Laplace and Goethe test problems. Error bars depict ± SEM, reflecting between-participant variance.

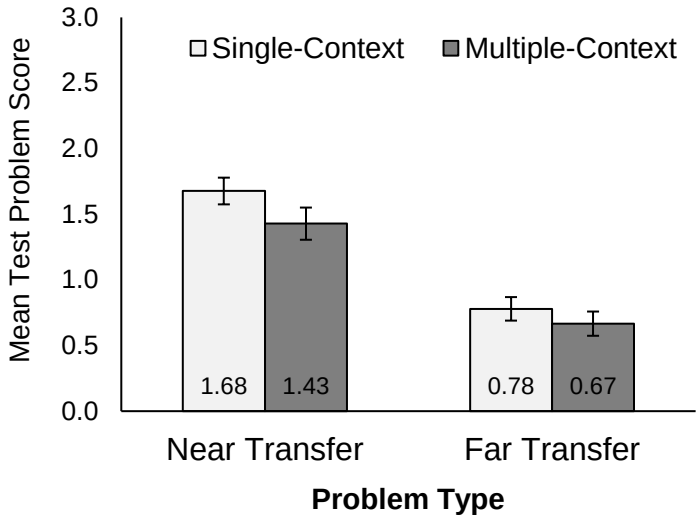


Figure 2. Test problem scores of Experiment 2 as a function of contextual variability (single-context, multiple-context) and test problem type (near transfer, far transfer). These scores reflect the quality of solution attempts for Laplace and Goethe test problems. Error bars depict ± SEM, reflecting between-participant variance.

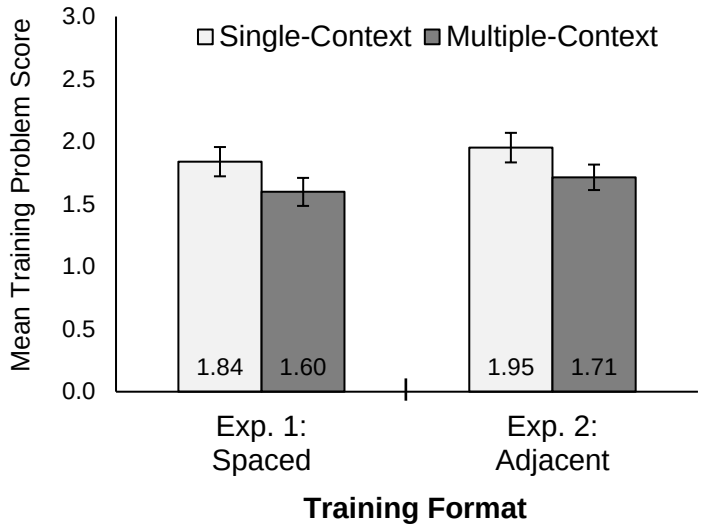


Figure 3. Training problem scores as a function of contextual variability (single-context, multiple-context) and training format (spaced, adjacent). These scores reflect the quality of solution attempts for Laplace and Goethe training problems. Error bars depict ± SEM.

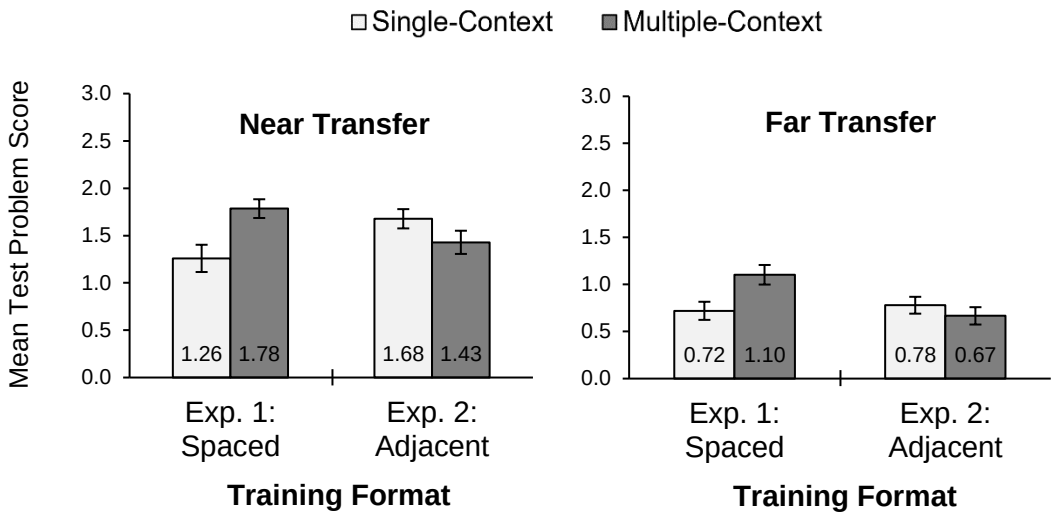


Figure 4. Data from the combined analysis of Experiments 1 and 2, with near and far transfer test performance as a function of contextual variability (single-context, multiple-context) and training format (spaced, adjacent). These scores reflect the quality of solution attempts for Laplace and Goethe test problems. Error bars depict ± SEM, reflecting between-participant variance.

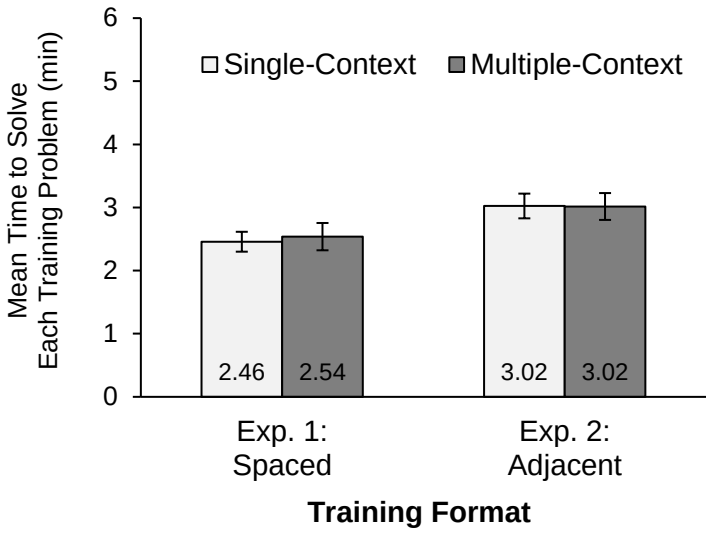
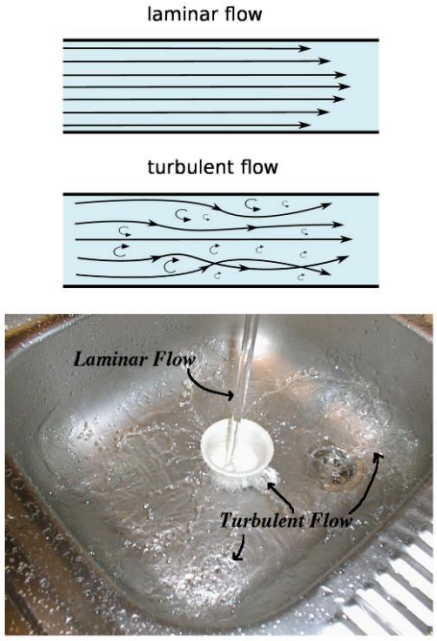


Figure 5. Mean time taken to solve each training problem in minutes as a function of contextual variability (single-context, multiple-context) and training format (spaced, adjacent). Error bars depict ± SEM.

Appendix A

Air flows into your lungs through the trachea (i.e., windpipe) and bronchioles, then eventually into alveoli where oxygenation of blood occurs. The air normally travels through the trachea and bronchioles in discrete and separate layers; the air molecules do not deviate from these layers or bump into each other. These layers are called 'laminae', hence the term laminar flow. The flow is unobstructed and generally silent. Irregularities in the flow such as a change in velocity, narrowing of the bronchioles, a thick or rough mucus lining around the passageways, or in some cases extremely 'thick' viscous air (e.g., smog), cause the air molecules to collide and disrupt the laminae. The flow of air therefore changes from laminar to turbulent, often accompanied by noise and a decrease in flow rate. Any pathology that causes narrowing or obstruction of an airway can produce turbulent flow.



There are many examples of laminar and turbulent flow in other parts of the body and in the outside world. For example, in an old house, calcium and other mineral deposits can build up in the copper pipes. Water can be heard gurgling through the pipe, and the pressure of the water coming out of the faucet will be reduced. Similarly, cigarette smoke rising in still air is initially laminar, but it becomes turbulent as it rises and cools (though you cannot hear it in this case).

Figure 6. The passage used to describe the principle of turbulent flow (labelled Goethe's law for the purpose of the experiments) during the training phase.

A patient with previous history of respiratory problems and asthma arrives at the ER wheezing and complaining of difficulty breathing. Investigation reveals excess mucous in the bronchioles and upper airway, thereby limiting the amount of air the patient can inhale and exhale. Explain why the mucous may have caused the patient's symptoms.

Figure 7. A sample training problem for Goethe's law in the context of the respiratory system.

Chapter 3: Do self-reported learning approaches predict transfer, and just how context specific are they?

Switching to the correlational approach, the goal of following line of research was to assess the predictive validity of the SPQ using performance measures that more precisely reflect transfer of learning. Given mixed findings in the literature, a secondary goal was to examine the degree to which SPQ scores tap into a stable, trait-like construct that generalizes across different learning contexts.

To examine whether SPQ scores reflect a trait-like construct, Study 1 involved a laboratory learning task involving three physiological principles that concluded with a final measure of transfer. Participants were then asked to complete the SPQ based on how they generally study. We reasoned that if SPQ scores reflect study approaches that are relatively stable within individuals, then they ought to predict transfer performance during this controlled laboratory task.

In Study 2 we then administered the SPQ during a university engineering course that focused on the application of various principles and equations. Students were asked to complete the SPQ based on their approaches for only this course, thus constituting a more context-specific implementation. We then correlated these SPQ scores with their performance on transfer-oriented exam questions identified by the instructor. We reasoned that this context-specific implementation would provide the most favourable conditions to detect the predicted positive correlation with SPQ deep approach scores. Further, a subset of students completed the SPQ for a different course that operated concurrently, allowing us to conduct a reliability analysis to estimate the degree to which SPQ scores ought to generalize across courses.

Do self-reported learning approaches predict transfer, and just how context-specific are they?

Andrew B. LoGiudice[1], Pakeezah Saadat[2], Julie Vale[3], Ryan Clemmer[3], Karen Gordon[3], Arshad Ahmad[4], Janette Barrington[5], Robert Cassidy[5], Sandra Monteiro[6], and Geoffrey R. Norman[6]

[1]McMaster University; Department of Psychology, Neuroscience & Behaviour,

[2]University of Toronto; Institute of Health Policy, Management, and Evaluation

[3]University of Guelph, School of Engineering

[4]Lahore University of Management Sciences; Office of the Vice Chancellor

[5]Concordia University; Department of Education, Centre for Teaching and Learning

[6]McMaster University; Health Research Methods, Evidence, and Impact

This research was supported in part by a SSHRC Canada Graduate Scholarship (Doctoral) awarded to the first author. Correspondence should be addressed to Andrew B. LoGiudice, Department of Psychology, Neuroscience & Behaviour, McMaster University, 1280 Main Street West, Hamilton, ON L8S-4K1. Email: logiudab@gmail.com

Abstract

As educators our goal is to equip students with flexible knowledge, allowing them to apply what they learn to new contexts that extend far beyond those encountered in the classroom. The same ideology is reflected by pervasive usage of the term 'deep learning' among educators and researchers. This desire for deep learning has led researchers to quantify it in educational settings. In particular, a self-report inventory called the Study Process Questionnaire (SPQ) is now widely administered to measure how much students engage in deep learning within a given learning environment, and hence to evaluate course design. The predictive validity of the SPQ remains uncertain, however, because studies seeking correlations between SPQ scores and student performance often do not make explicit the nature of the performance measures being adopted. Further obscuring the SPQ's predictive validity, it remains unclear whether SPQ scores obtained from students in one learning environment can be generalized to other learning environments. Here we tackled these two measurement issues with emphasis on performance measures that reflect transfer of learning to new contexts. We found that SPQ scores did not predict performance in a laboratory transfer task, nor did they predict performance on transfer-oriented exam questions within a university engineering course. Analysis of SPQ scores obtained from the same students in different courses also showed that the constructs tapped by the SPQ ought to be relatively trait-like and generalize across learning contexts. Taken together, these data call into question the utility of the SPQ for improving educational practice.

Keywords: Approaches to Learning, Study Process Questionnaire, Transfer of Learning

Introduction

Deep learning

Education, at its core, is the pursuit of generalizable knowledge. We often begin by asking students to remember individual definitions, facts, and examples, but the true goal is to have them grasp higher-order structures—the general laws, equations, principles, and so on—that tie all the information together. Thus, our hope is that students will begin to extrapolate beyond the specifics and understand the deeper principles at play, granting them greater flexibility in how they later apply these far-reaching principles to novel situations.

The same idea lies at the heart of the phrase 'deep learning' that now pervades educational research and practice. Although deep learning has no universally accepted definition, it is usually described in terms of higher-order learning (e.g., Agarwal, 2019; Bloom, 1956; Krathwohl, 2002), critical thinking (e.g., Kek & Huijser, 2011; Laird et al., 2014; Phan, 2011), the understanding of key principles (e.g., Biggs, 1987; Kember et al., 1997; Laird et al., 2008), and the application of knowledge to new contexts (e.g., Chernobilsky et al., 2004; Halpern & Hakel, 2003). These descriptions make it clear that deep learning ought to produce students who can generalize their understanding to solve novel problems.

The SPQ

As a testament to how popular this deep learning terminology has become, a ubiquitous measurement tool called the SPQ (i.e., Study Process Questionnaire; Biggs et al., 2001) has been trumpeted as a way to measure how much students adopt "deep learning approaches". Briefly, the SPQ prompts students to respond to Likert items regarding their strategies or attitudes within a given educational context, producing a deep approach score and surface approach score for each student. The deep approach score is said to reflect a student's tendency to process content "at a high level of generality, such as main ideas, themes, and principles", whereas the surface approach score is said to reflect a tendency to engage in rote memorization or other superficial forms of processing "extrinsic to the real purpose of the task" (Biggs, 1993, p. 7). Put simply, students who receive higher SPQ deep approach scores ought to engage in more deep learning, presumably resulting in generalizable knowledge that can be applied to many new contexts.

The SPQ is certainly enticing from a practical perspective. Regardless of the educator's goals, much of learning is governed by what the student actually chooses to do (e.g., Biggs,

1999), and so the SPQ might help educators gauge whether students are likely to adopt deep learning for educational contexts where learning is largely self-directed. The SPQ also seems like a reasonable way to evaluate whether specific educational interventions increase deep learning. Speaking to this latter point—and illustrating how directly this tool continues to shape educational practice—many authors continue to endorse educational approaches based on corresponding increases in SPQ deep approach scores (for review, see Baeten et al., 2010).

Do SPQ scores lack predictive validity?

But some authors have recently criticized the SPQ for a lack empirical evidence to support its predictive validity (Dinsmore & Alexander, 2012; Howie & Bagnall, 2013). Most notably, if the SPQ quantifies a student's tendency to adopt deep learning, one might expect its deep approach scores to positively correlate with aggregate measures of academic achievement like GPA or performance on course assessments. Some studies do report a small positive correlation in the expected direction (e.g., Cantwell & Moore, 1998; McManus et al. 1998; Scouller & Prosser, 1994; Snelgrove & Slater, 2003; Zeegers, 2001) but many others do not (e.g., Gijbels et al., 2005; Groves, 2005; Wilding & Andrews, 2006). Amplifying these criticisms, a large meta-analysis found that SPQ deep approach scores positively but very weakly correlated with academic achievement (r^{+} = .14), accounting for less than 3% of variance (Richardson et al., 2012). Such modest findings suggest the SPQ may not capture the deep learning construct as it is meant to.

Two unresolved measurement issues

Our goal was to further assess the predictive validity of the SPQ by tackling two unresolved measurement issues. First, we discuss how the SPQ's ability to predict academic achievement in past studies is obscured by heterogeneous and ill-defined outcomes measures, suggesting that more precisely defined outcome measures are needed. Second, we discuss how the SPQ's ability to predict academic achievement is obscured by its alleged context specificity, meaning it remains unclear whether the SPQ measures transient states of the student that depend on the specific learning context, or stable traits that are likely to persist across various learning contexts. We then describe two studies meant to address both issues.

Issue #1: How well do SPQ scores predict transfer?

One explanation for the modest predictive power of the SPQ in past studies is that they have largely relied on aggregate performance measures that are highly heterogenous, and so might not have indexed generalization of knowledge per se (Watkins, 2001, as cited in Choy et al., 2012). For example, a student who takes exclusively memorization-oriented courses might adopt a surface approach and yet perform well on final examinations, in which case exam performance and GPA are poor indicators of deep learning. This caveat suggests that more precise outcomes measures are needed to properly test the SPQ's predictive validity.

Here we suggest the SPQ's predictive validity rests foremost on how well it predicts transfer of learning. Transfer researchers are interested in how a learner abstracts general principles from examples based on the deep structural relations that are held constant. Whereas an expert's success at transferring a principle to solve a new problem is largely independent of the problem's particular "surface features"—i.e., superficial aspects of the problem not inherent to the principle—a novice's success is highly dependent on the degree to which the problem's surface features match those of previously seen examples (e.g., Chi et al., 1981; Holyoak & Koh, 1987; Ross, 1987; for reviews, see Day & Goldstone, 2012; Reeves & Weisberg, 1994). By this logic, an apt measure of deep learning is the ability for learners to solve problems instantiated in novel contexts. We will return to this point shortly.

Issue #2: How context-specific are SPQ scores?

Another consideration obscuring the SPQ's predictive validity is its reputed context specificity. SPQ scores obtained from students are argued to be highly sensitive to the specifics of a given learning context, including the content being learned, the instructor's style of assessment, prior knowledge, and the student's perceptions of learning objectives (e.g., Biggs, 1993; Baeten et al. 2010; Struyven et al. 2006; Wilson & Fowler, 2005). Following this logic, studies have reported substantial gains in SPQ scores after implementing an educational intervention and suggest that these findings support the SPQ's context specificity (e.g., Cowman, 1998; Kember et al., 1997; Nijhuis et al., 2005). Such findings have also led authors to state that it is "inappropriate to categorize students as 'surface' or 'deep' learners on the basis of SPQ responses . . . as if an approach score measured a stable trait of the individual" (Biggs et al., 2001, p. 136).

However, it remains unclear whether SPQ scores reflect stable individual differences. For example, one study found substantial overlap between learning approaches and personality traits (Duff et al., 2004) whereas others found that learning approaches and personality traits are related but account for enough unique variance to be considered distinct constructs (Chamorro-Premuzic & Furnham, 2009; Chamorro-Premuzic et al., 2007; Zhang, 2003). Another study has also reported a substantial association between learning approaches and "cognitive styles" that denote stable individual preferences for information processing (Bouckenooghe et al., 2016). Lastly, at face value, many SPQ items (e.g., "I feel that virtually any topic can be highly interesting once I get into it") seem to describe a student's perceptions of their learning *in general* rather than for a specific learning context. In short, the context specificity of the SPQ requires further study.

Note how the context specificity of the SPQ has important theoretical and practical implications. If the SPQ is indeed highly context-specific, educators should be cautious about extrapolating scores beyond a single learning context. This would also suggest that aggregate outcome measures that span multiple courses (e.g., GPA) may not paint an accurate picture of the SPQ's predictive validity. Conversely, if the SPQ captures stable individual differences in strategies or attitudes that persist across various learning contexts, then educators could use it as a remedial tool to identify students who are not adopting deep approaches early within a larger curriculum. Either conclusion affects how the SPQ ought to be studied and used in practice.

Present Studies

To tackle the first issue of ill-defined learning outcomes, we sought to define our outcome measures more precisely through the lens of transfer. Study 1 was a laboratory study in which we had participants learn about several basic science principles, then assessed how well their SPQ scores predicted their ability to generalize the principles to solve new problems. Next, Study 2 was a classroom study in which we had engineering students complete the SPQ during their course, then selectively examined performance on examination questions that the instructor identified as demanding transfer. If the SPQ captures a student's tendency to engage in deep learning, and deep learning gives rise to transfer, then deep approach scores obtained from participants ought to correlate positively with their transfer performance.

To tackle the second issue of context specificity, we sought to compare the SPQ's predictive power when completed in a context-general or context-specific manner. In Study 1 participants were asked to complete the SPQ based on their general study behaviours without a specific course in mind. In Study 2 participants were instead asked to complete the SPQ based on their experiences in the two specific engineering courses under investigation. If SPQ scores are indeed highly context-specific, then they ought to better predict a student's transfer performance when obtained in relation to a relevant course (i.e., Study 2) than when they complete it without a specific course in mind (i.e., Study 1). Moreover, some students completed the SPQ within two separate engineering courses for Study 2, allowing us to estimate the extent to which deep approach scores obtained from students in one course will generalize to other courses.

Study 1: Context-General Use of the SPQ

The goal of Study 1 was to determine whether SPQ scores predicted transfer in a context-general manner during a controlled laboratory exercise. We had undergraduate participants study three basic science principles in a self-paced manner, take a final transfer test that assessed their ability to generalize the principles to solve new problems, then complete the SPQ based on how they perceive their learning approaches in general—i.e., without any reference to a particular course or educational context.

Method

Participants

A total of 221 undergraduate students (Mean Age = 19.0, SD = 3.3; 79% female) enrolled in an introductory psychology course at McMaster University participated for course credit. The study was approved by the McMaster Research Ethics Board (Protocol #1955), and all participants gave informed consent.

Materials

We adopted similar learning materials as those used in prior transfer studies (Kulasegaram et al., 2012; Kulasegaram et al., 2017). Briefly, these materials focus on three physiology principles: *Laplace's Law* describes the relation between radius, pressure, and tension in a cylindrical vessel; turbulent flow, hereafter labeled *Goethe's Law* for consistency, describes the relation between flow of fluid in a vessel and fluid velocity, fluid viscosity, vessel

radius, and vessel roughness; lastly, *Starling's Law* describes the relation between ventricular stretching and stroke volume in the heart.

For each of these three principles we used: (i) a training passage that explained how the principle applies to a specific organ system; (ii) three training problems that each outlined a clinical scenario in the context of a specific organ system, prompting the participant to explain how the principle applies; and (iii) several test problems that each outlined clinical scenarios in the context of a specific organ system, prompting the participant to choose which of the three principles best applied and why.

We also manipulated whether test problems were embedded in the same organ system as a relevant prior training problem (i.e., near transfer) or a novel organ system context that had not been seen during the experiment (i.e., far transfer). Altogether there were seven near transfer test cases (2 Laplace, 2 Goethe, 3 Starling) and six far transfer test cases (3 Laplace, 3 Goethe).

Procedure

Participants were seated at individual computer booths in a quiet laboratory setting. Learning materials were shown on a computer monitor and all written responses were typed using a standard keyboard. The study was self-paced with the restriction that participants could not go back to any particular training material or test question shown on the monitor once they had decided to proceed past it. The whole study took approximately 1–1.5 hours (Mean = 69.1 min, SD = 20.1) and was programmed using PsychoPy2 software (Peirce, 2007).

During the training phase, participants were asked to learn three basic science principles for an upcoming test where they would have to generalize the principles to solve new problems. For each principle they began by studying its associated training passage and then pressed a key to proceed to three training problems. The first training problem appeared and prompted the participant to explain how the principle they just read about applied to the problem. Once the participant pressed a key to proceed, their typed response was replaced by normative written feedback that explained the correct answer. The participant then pressed a key to proceed to the next training problem. This same process repeated for three training problems. Once all three training problems were completed for a given principle, training on the next principle began. The order of principles and training problems for each principle was randomized separately for each

participant. In sum, for each of the three principles, participants studied a training passage and attempted to solve a series of training problems with immediate feedback.

During the subsequent test phase, participants were asked to solve a series of 13 test problems without any feedback. Order of test problems was randomized for each participant. For each test problem the participant was asked to identify which of the three principles best applied and to type a brief explanation of their reasoning.

Lastly, participants completed the R-SPQ-2F. Each item was shown in sequence on the monitor. Participants could not go back to an item after submitting a response. Critically, they were asked to respond based on how they study *in general*—not how they studied the three principles during this study. The critical question was whether these SPQ scores obtained in a context-general manner could predict their transfer performance.

Results

Scoring

Solution attempts for training and test problems were all scored from 0–3 with a higher score reflecting greater understanding of the correct principle (Kulasegaram et al., 2017). The scoring scheme was as follows: 0 = Does not capture correct principle; 1 = Captures correct principle but with obviously flawed understanding; 2 = Captures correct principle but with key ideas absent or misconstrued; 3 = Captures correct principle with all key ideas present and described accurately.

To ensure reliability in scoring, solution attempts for test problems from a subset of 55 participants were scored by both the lead author and an independent rater. Inter-rater reliability was high for both near transfer problems (ICC = .90) and far transfer problems (ICC = .89), so all test problem solution attempts were scored by the lead author in a blinded fashion.

Exclusions

Data from 7 participants were excluded because they left more than four test problem responses blank, suggesting noncompliance. Data from 3 participants were also excluded due to software failure. This left 211 participants in the following analyses.

Data Quality

Internal consistency of SPQ deep approach scores (α = .77) and surface approach scores (α = .72) was adequate, suggesting that the various items of each scale were tapping into the same construct.

Transfer performance was measured by collapsing across performance for near and far transfer problems to achieve adequate reliability. Internal consistency of the resulting transfer measure was high (α = .70), suggesting it was suitable for detecting the predicted positive correlation with SPQ deep approach scores.

Correlating SPQ Scores and Transfer Performance

As shown in Figure 1, the correlation[9] between SPQ deep approach scores and transfer performance was not significant and very close to zero (r = .01, p = .88). The same was true of the correlation between SPQ surface approach scores and transfer performance (r = .03, p = .70).

< Figure 1 here >

Discussion

SPQ scores obtained in a context-general manner clearly did not predict transfer of knowledge during this laboratory task. However, we emphasize that these results must be interpreted with caution because they constitute absence of evidence.

One advantage of this study relative to prior studies is that it more precisely defines deep learning as the ability to transfer knowledge to analogous problems. Indeed, a similar learning task has been used in several prior studies examining transfer (Kulasegaram et al., 2012, 2015, 2017), raising confidence that the transfer construct is sound. Another strength is that participants were from a range of disciplines in the sciences, social sciences, and humanities. Our sample is therefore representative of populations of interest to SPQ researchers.

Perhaps the most salient limitation of this study is that we used the SPQ in a way it was not intended to be used. Specifically, the SPQ was designed to be implemented in authentic educational settings where—unlike the contrived laboratory setting used here—many real-world

[9] These data are from four studies using the same general design, differing only in instructions, training problems, or the format of presenting training problems. We collapsed across these independent variables because they do not pertain to the research question addressed here, and because the correlations of interest were comparable across studies. See Appendix A for details.

variables (e.g., motivating factors, educator's assessment style, perceptions of learning objectives) are likely to interact and affect responses to the SPQ (e.g., Biggs, 1993; Baeten et al. 2010; Struyven et al. 2006; Wilson & Fowler, 2005). Given the present null findings, this leaves one to wonder: will the SPQ better predict transfer performance when completed in a context-specific manner? This question was the focus of Study 2.

Study 2: Context-Specific Use of the SPQ

In Study 2 we administered the SPQ during two undergraduate-level engineering courses. For one of these courses we examined performance on a subset of assessments that the instructor perceived as tapping into transfer[10]. We reasoned that this context-specific use of the SPQ paired with our transfer-oriented outcome measures would provide more favourable conditions to detect the predicted positive correlation between SPQ deep approach scores and transfer. In addition, because some students took both courses, we also performed a reliability analysis to infer how much the obtained SPQ scores would be expected generalize across courses.

Method

Participants

A total of 308 undergraduate students enrolled in an engineering systems analysis course at the University of Guelph were invited to participate for course credit. Similarly, a total of 228 undergraduate students enrolled in a material science course at the same university were invited to participate for course credit. As noted earlier, some of these students were enrolled in both courses. The study was approved by the University of Guelph Research Ethics Board (Protocol #17-07-007), and all participants included in the final analysis gave informed consent.

Briefly, the systems analysis course focused on the integration and extension of concepts or equations that students had learned in prerequisite courses. Instruction therefore centered around application of concepts to new contexts rather than the initial learning of those concepts. In contrast, the material science course was more focused on the memorization of definitions and the properties of various materials.

[10] We originally planned to examine correlations between SPQ scores and examination performance for both courses. However, the exam scores for the material science course had very low internal consistency (), so we will not report the corresponding correlations with SPQ scores here.

Procedure

Students were notified about the study near the end of their respective courses and received a link to an online survey. The survey asked students for basic demographic information; to respond to each of the SPQ items based on their study approaches for the course being targeted; and for permission to analyze their survey responses in relation to their performance on course assessments.

The instructor of the engineering systems analysis course, one of the authors of this paper (JV), identified a subset of ten midterm and final examination questions that they perceived as tapping into memorization and a separate subset of eight questions that they perceived as tapping into transfer. Memorization questions were classified on the basis of being highly similar to examples or problems that students had seen previously during the course (e.g., posing the same problem but with different values in the equations). Conversely, transfer questions were classified on the basis of: (i) being dissimilar to any previously encountered examples or problems; (ii) providing an abundance of information, some of which being irrelevant to solving the problem; and (iii) open-endedness in that students could adopt of many approaches to arrive at the correct answer. Sample questions of each type are shown in Figure 2, and a schematic of the overall study design is shown in Figure 3.

< Figure 2 here >

< Figure 3 here >

Results

Response Rate

A total of 131 students (48% female) from the systems analysis course completed the online survey and consented to having their data included in the study, indicating a 42.5% response rate. Of these students, 45 (44% female) also completed the survey for the concurrent material science course.

Exclusions

Seven of the students in the systems analysis course were dropped from analysis because they did not complete one of the course assessments, leaving 124 students in the final analysis.

Data Quality

For the systems analysis course, internal consistency was low for memorization performance ($\alpha = .54$) but good for transfer performance ($\alpha = .70$). This discrepancy is likely explained by restriction of range, as the standard deviation of the memorization measure (SD = 9.6) was much lower than that of the transfer measure (SD = 23.2). We will still report correlations between SPQ scores and both memorization and transfer measures for archival purposes, but we emphasize that correlations with the memorization measure should be interpreted with caution due to its low reliability.

Correlating SPQ Scores and Test Performance

Four correlations are shown in Figure 4. As seen in the upper panels, the correlation between SPQ deep approach scores and transfer performance was not significant ($r = .07$, $p = .46$). The same was true of the correlation between SPQ surface approach scores and transfer performance ($r = -.08$, $p = .37$). As seen in the bottom panels, the correlation between SPQ deep approach scores and memorization performance was larger but not significant ($r = .15$, $p = .09$). Likewise, the correlation between SPQ surface approach scores and memorization performance was larger but not significant ($r = -.17$, $p = .06$). Although this latter correlation approached significance, note how it is in the opposite direction of what would be predicted; a high surface approach score ought to be *positively* correlated with performance on memorization questions. In sum, these data do not support the SPQ's predictive validity.

< Figure 4 here >

Reliability Analyses

Because 45 students completed the SPQ for both courses, we also performed reliability analyses to infer how well SPQ scores obtained from students would be expected to generalize to other courses. An intraclass correlation was calculated on the SPQ deep approach scores from both courses as follows: the numerator consisted of variance accounted for by student; and the denominator consisted of variance accounted for by student, variance accounted for by course, and error variance. This analysis yielded an intraclass correlation of .64, suggesting moderate generalization of deep approach scores across courses. An identical analysis was then conducted on the surface approach scores. This analysis yielded an intraclass correlation of .45, suggesting moderate generalization of SPQ surface approach scores across courses, albeit less than for deep approach scores.

Discussion

SPQ scores obtained in a context-specific manner did not predict performance on assessments that the instructor viewed as tapping into transfer. As for our question of context specificity, the reliability analyses indicated that SPQ scores from students should generalize to a reasonable degree across courses, particularly the deep approach scores.

A strength of this study is its ecological validity because the data were collected in an authentic educational setting with minimal changes to the way students were taught or tested. SPQ scores were also obtained in relation to a specific course, ruling out the interpretation that we found no correlations simply because the tool was not used in a context-specific manner.

A potential weakness was the method for classifying test questions. Although the instructor (JV) felt it was relatively intuitive to classify examination questions as tapping either memorization or transfer, their subjective perceptions of transfer may not align with the notion of generalizable knowledge as discussed by researchers. This interpretation seems unlikely, however, because the instructor has taught the course content for over ten years, is actively engaged in research on problem solving, and also because the format of the systems analysis course (i.e., focusing on integration and application of broad principles and equations) naturally lends itself to the transfer of knowledge.

General Discussion

Our two goals here were to assess the predictive validity of the SPQ using transfer-oriented outcome measures, and to examine the context specificity of SPQ scores. We did not find significant correlations between SPQ approach scores and transfer performance in a laboratory setting when student learning approaches were measured in a trait-like manner. We also did not find significant correlations in an authentic classroom setting when student learning approaches were measured in relation to a specific course they were taking. Thus, the SPQ did not predict transfer of knowledge in either a context-general or context-specific implementation.

We interpret these results by suggesting that the SPQ may not measure what it is intended to measure. Specifically, SPQ scores are based solely on the self-reports of students, and so may reflect a student's *perceptions* of their approaches rather than the approaches they actually adopt in practice (also see Veenman et al., 2003). This interpretation also seems plausible given that most studies examine correlations between SPQ scores and aggregate performance measures,

with little discussion about the specific mental processes that characterize deep learning or their associated outcomes. In short, the SPQ sporadically predicts aggregate learning outcomes, but it may be doing so without capturing the so-called deep learning behaviours it is meant to index.

With respect to our secondary question about context specificity, the reliability analysis in Study 2 showed that SPQ deep approach scores ought to generalize a fair amount across different courses. This result seems somewhat surprising given that many authors argue that SPQ scores are highly specific to the learning context in which they were obtained (e.g., Biggs, 1993; Baeten et al. 2010; Struyven et al. 2006; Wilson & Fowler, 2005). Thus, these findings support the more nuanced view that SPQ scores partially reflect context-general behaviours that persist across courses *and* context-specific behaviours that hinge upon the particular course under investigation. Perhaps the real conclusion, then, is that the SPQ is more context-general than it is often perceived to be, especially in light of observed correlations between SPQ scores and other stable individual differences in personality (Duff et al., 2004; Chamorro-Premuzic & Furnham, 2009; Chamorro-Premuzic et al., 2007; Zhang, 2003) and information processing (Bouckenooghe et al., 2016). Regardless, discussions about the context specificity of the SPQ are arguably rendered moot if it does not predict learning outcomes, as was the case in the two studies described here.

It is worth noting that another context-general tool has shown greater promise as a way to predict transfer via stable individual differences (DeLosh et al., 1997). In essence, the tool is a simple function-learning task in which participants first receive several training examples (i.e., *x* values and *y* values from a bi-linear "V" function) and then try to predict the *y* values that correspond to new *x* values. Studies adopting this function-learning task have found clear dissociations in that some participants (labelled "abstraction learners") reliably provide *y* values matching the bi-linear function whereas others (labelled "exemplar learners") do not, suggesting difference in how participants processed the training examples (Frey et al., 2017; McDaniel et al., 2014; McDaniel et al., 2018). The two learner profiles also predicted performance on two other basic categorization tasks (McDaniel et al., 2014, Experiments 1c, 2) and student performance on application questions, but not memorization questions, within real university chemistry courses (McDaniel et al., 2018). Such findings support the stability and generalizability of this observed individual difference in authentic educational settings.

Also note how this function-learning task focuses on performance rather than introspection. Self-report inventories like the SPQ rely on student perceptions of their motivation and learning strategies, thus raising concerns about inaccurate introspection (e.g., Dunlosky & Rawson, 2012; Nisbett & Wilson, 1977). An informal glance at the SPQ's items also reveals that most pertain to general motivation as opposed to specific forms of information processing (e.g., "I find that I have to do enough work on a topic so that I can form my own conclusions before I am satisfied"; "I only study seriously what's given out in class or in course outlines"). In contrast, the function-learning task lends itself to discussion of particular forms of information processing that promote subsequent transfer (e.g., whether learners are focusing on each training example in isolation or commonalities between examples). Proponents of the SPQ therefore seem to take a generalist stance—i.e., students who have more motivation will *generally* learn more—whereas the function-learning task emphasizes the need to understand *specific* mental operations enacted during training. This point is critical because learners seldom adopt transfer-conducive strategies without explicit instructions (e.g., Gentner et al., 2003; Gick & Holyoak, 1983; for review, see Alfieri et al., 2013). Therefore, using motivation to predict transfer outcomes might be painting the problem with too broad a brush, overlooking the connection between specific mental processes and their outcomes.

Conclusion

We end on a cautionary note. Given the vast uptake of the SPQ in educational circles, it is vital to test its predictive validity using more precisely defined outcome measures. The present study did this by examining the association between SPQ approach scores and outcome measures that were meant to selectively tap transfer of knowledge. Nonetheless, the null findings we report here further challenge the construct of deep learning as measured by the SPQ. We encourage future work to move in a similar direction by avoiding overly broad definitions of deep learning, instead focusing on specific mental processes and how they are linked to specific learning outcomes. Until more empirical evidence is provided, however, we argue that the current evidence is insufficient to warrant use of the SPQ for its intended purpose.

Acknowledgements

This research was supported in part by a SSHRC Canada Graduate Scholarship (Doctoral) awarded to the first author. We thank attendees of the McMaster Conference on Education & Cognition, 2018, for insightful questions that contributed to this line of research.

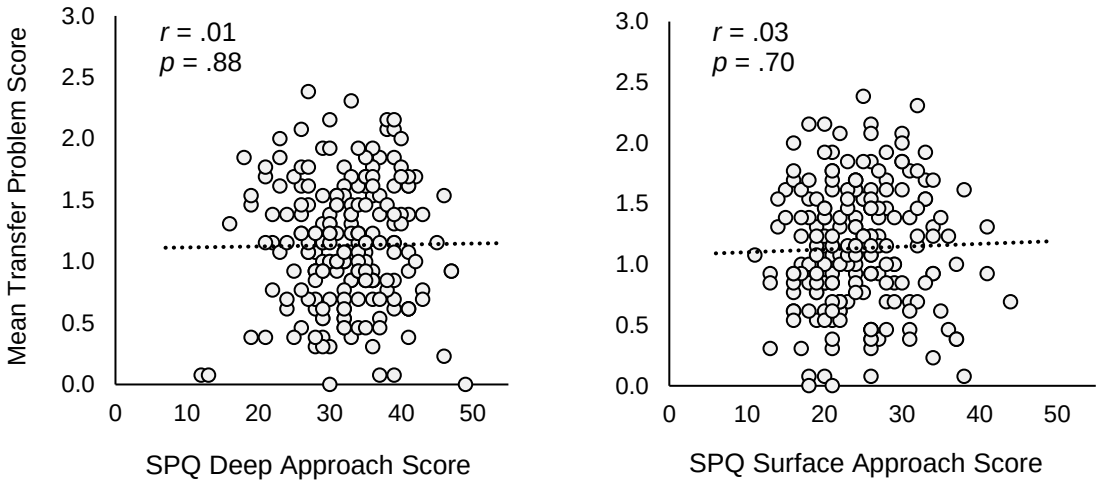


Figure 1. Correlations between SPQ approach scores (Deep, Surface) and transfer performance in Study 1 (N = 211).

Transfer Question

You arrive home after a long day of studying and decide to bake a cheese pizza. You preheat the oven for 10 minutes to a temperature of 220 °C and cook the pizza for 25 minutes. Immediately after removing the pizza from the oven, the pizza is at a temperature of 85 °C. The pizza has a diameter of 34 cm. The cheese is approximately 2 cm thick and the crust is 3 cm thick. (Assume the cheese continues to the edge of the crust).

Cheese has a specific heat capacity of 2.8 kJ/kg °C and a density of 1050 kg/m^3. The total resistance of the crust is estimated to be 6 °C/W and the convective coefficient of the air above the pizza is 100 W/m^2K. Assume an ambient temperature of 20 °C and that the capacitance of the crust is negligible.

You want to know how long you must wait to consume your pizza. Use any method you wish to find the differential equation that you need to solve this problem. You must show your work for full points, including the calculation of any relevant R and C parameters.

Memorization Question

A particular system has input x, output y and differential equation $2\dot{y}(t)+ 4y(t) = 3x(t)$

If the system has input $x(t) = 7u(t)$ and initial condition $y(0^-) = -6$, write the equation for the complete response. Indicate the free and the forced portions of your solution.

Figure 2. Sample transfer and memorization questions from the engineering systems analysis course in Study 2. Note that the transfer problem on the left was dissimilar to any examples or problems encountered previously in the course, contained information irrelevant to the solution, and could be solved several different ways based on a broad set of principles and equations that students had learned. In contrast, the memorization question on the right was the same as a previous problem encountered during the course but with different values in the equations.

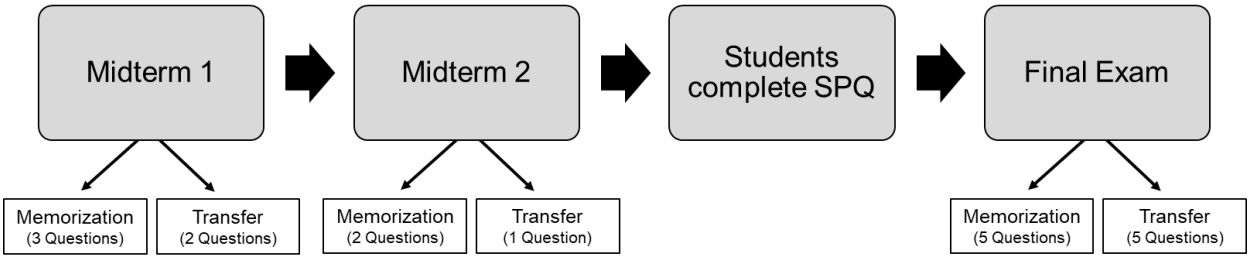


Figure 3. A schematic of the design used for the engineering systems analysis course in Study 2.

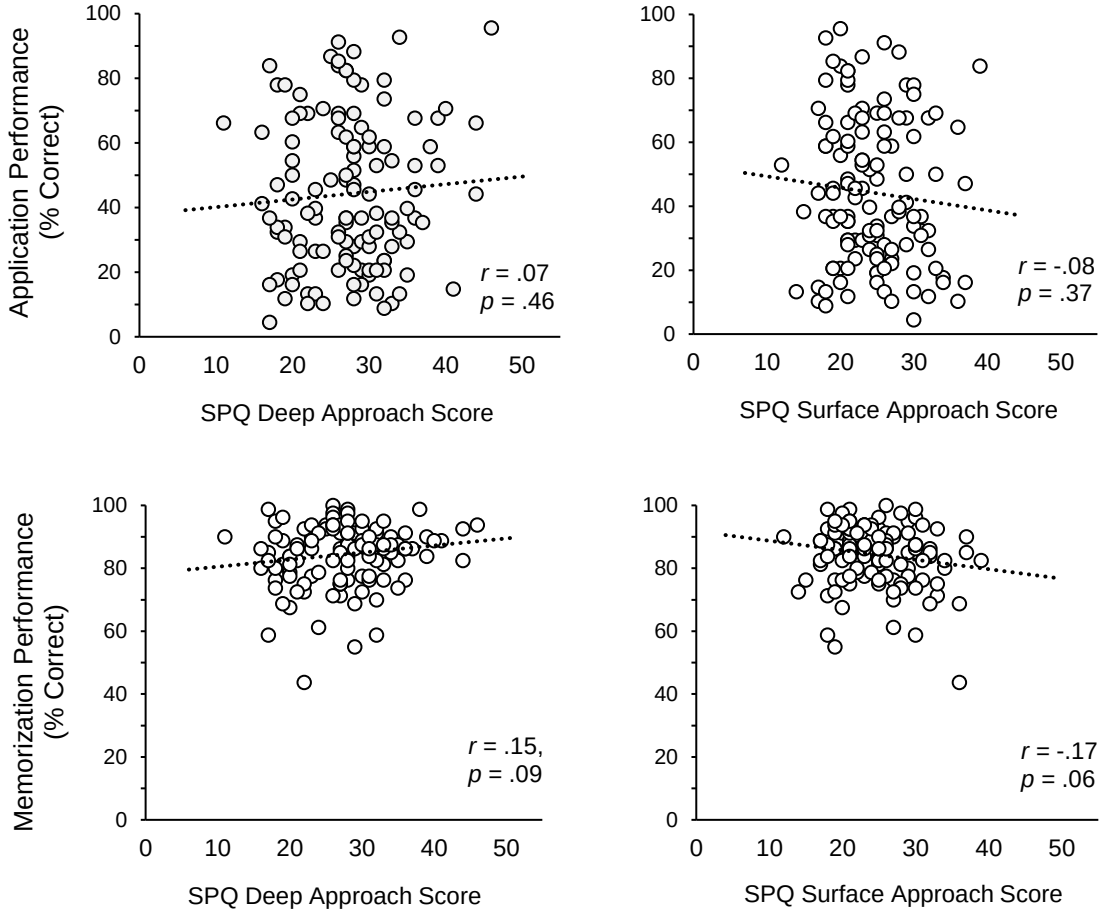


Figure 4. Correlations between SPQ approach scores (Deep, Surface) and test question performance (Transfer, Memorization) in Study 2 (N = 124).

Appendix A

		Near Transfer		Far Transfer	
Experiment	***N***	**Surface**	**Deep**	**Surface**	**Deep**
1	57	+ .25, p = .07	– .13, p = .33	+ .05, p = .69	– .02, p = .91
2	59	– .12, p = .37	+ .02, p = .87	+ .08, p = .53	+ .02, p = .87
3	38	– .08, p = .63	– .09, p = .59	+ .20, p = .23	– .16, p = .34
4	57	– .15, p = .27	+ .17, p = .20	– .06, p = .67	+ .15, p = .26

Table 1. Pearson *r* values and corresponding *p* values for the four experiments that constituted the full data set of Study 1. None of these correlations were statistically significant.

Experiment	IV_1: Additional Instructions	IV_2: Training Problem Contexts	IV_3: Training Problem Presentation
1	None	1 Context vs. 3 Contexts	Sequential
2	None	1 Context vs. 3 Contexts	Adjacent
3	Generalize	1 Context vs. 3 Contexts	Adjacent
4	Associate	1 Context vs. 3 Contexts	Adjacent

Table 2. A summary of the three between-participant independent variables manipulated across the four experiments outlined above in Table 1. These experiments differed in three independent variables: (1) additional instructions during training; (2) specific contexts of training problems; and (3) training problem presentation. These manipulations are described in more detail below.

IV_1: Instructions

Just before studying the training passage for each principle, a set of instructions might be shown:

Generalize:
It is VERY IMPORTANT to keep in mind that basic medical principles, like those you are studying here, are often applicable in several different organ systems and a wide variety of other contexts throughout the human body.

So even though all the examples you will see in this experiment are associated with a specific organ system or context, try your best NOT to be distracted by these specific contexts, and to think about the underlying principles as general principles that are widely applicable throughout the human body.

Associate:
It is VERY IMPORTANT to keep in mind that basic medical principles, like those you are studying here, often apply only in one (or a very few) particular organ system or context in the human body.

The examples you will see in this experiment are each embedded in a specific organ system or context, so try your best to CLOSELY ASSOCIATE the underlying principles with the specific context(s) they are presented in.

IV_2: Context Variation at Training

Training problems for Laplace's and Goethe's Law were either in one or three organ system contexts:

One Context:
Laplace = 3 gastrointestinal
Goethe = 3 respiratory

Three Contexts:
Laplace = 1 gastrointestinal, 1 cardiovascular, 1 lymphatic
Goethe = 1 respiratory, 1 cardiovascular, 1 nervous

IV_3: Training Problem Presentation

Training problems were presented either adjacent to one another on the monitor or sequentially with a brief distractor task in between:

Adjacent:
For a given principle, any preceding training problems and their corrective feedback remained on the monitor while participants solved subsequent training problems.

Sequential:
For a given principle, participants completed a brief filler task (i.e., given a random category label and typed as many category members as possible in 45 seconds) after studying the training passage and in between training problems. Prior training problems and their corrective feedback were not present on the monitor while solving subsequent training problems.

References

Agarwal, P. K. (2019). Retrieval practice & Bloom's taxonomy: Do students need fact knowledge before higher order learning? *Journal of Educational Psychology, 111*(2), 189–209.

Alfieri, L., Nokes-Malach, T. J., & Schunn, C. D. (2013). Learning through case comparisons: A meta-analytic review. *Educational Psychologist, 48*(2), 87–113.

Baeten, M., Kyndt, E., Struyven, K., & Dochy, F. (2010). Using student-centred learning environments to stimulate deep approaches to learning: Factors encouraging or discouraging their effectiveness. *Educational Research Review*, *5*(3), 243–260.

Biggs, J. B. (1987). *Student Approaches to Learning and Studying. Research Monograph*. Australian Council for Educational Research Ltd., Radford House, Frederick St., Hawthorn 3122, Australia.

Biggs, J. (1993). What do inventories of students' learning processes really measure? A theoretical review and clarification. *British Journal of Educational Psychology*, *63*(1), 3–19.

Biggs, J., Kember, D., & Leung, D. Y. (2001). The revised two-factor study process questionnaire: R-SPQ-2F. *British Journal of Educational Psychology*, *71*(1), 133–149.

Bloom, B.S. (Ed.), Engelhart, M.D., Furst, E.J., Hill, W.H., & Krathwohl, D.R. (1956). *Taxonomy of educational objectives: The classification of educational goals*. Handbook 1: Cognitive domain. New York: David McKay.

Bouckenooghe, D., Cools, E., De Clercq, D., Vanderheyden, K., & Fatima, T. (2016). Exploring the impact of cognitive style profiles on different learning approaches: Empirical evidence for adopting a person-centered perspective. *Learning and Individual Differences*, *51*, 299–306.

Cantwell H. & Moore P.J. (1998) Relationships among control beliefs, approaches to learning and the academic achievement of final year nurses. *The Alberta Journal of Educational Research, 1*, 98–102.

Chamorro-Premuzic, T., & Furnham, A. (2009). Mainly Openness: The relationship between the Big Five personality traits and learning approaches. *Learning and Individual Differences*, *19*(4), 524–529.

Chamorro-Premuzic, T., Furnham, A., & Lewis, M. (2007). Personality and approaches to learning predict preference for different teaching methods. *Learning and Individual Differences*, *17*(3), 241–250.

Chernobilsky, E., DaCosta, M. C., & Hmelo-Silver, C. E. (2004). Learning to talk the educational psychology talk through a problem-based course. *Instructional Science*, *32*(4), 319–356.

Chi, M. T. H. H., Feltovich, P. J., & Glaser, R. (1981). Catagorization and representation of physics problems by experts and novices. *Cognitive Science*, *5*(2), 121–152.

Choy, J. L. F., O'Grady, G., & Rotgans, J. I. (2012). Is the Study Process Questionnaire (SPQ) a good predictor of academic achievement? Examining the mediating role of achievement-related classroom behaviours. *Instructional Science*, *40*(1), 159–172.

Day, S. B., & Goldstone, R. L. (2012). The import of knowledge export: Connecting findings and theories of transfer of learning. *Educational Psychologist*, *47*(3), 153–176.

Dinsmore, D. L., & Alexander, P. A. (2012). A critical discussion of deep and surface processing: What it means, how it is measured, the role of context, and model specification. *Educational Psychology Review*, *24*(4), 499–567.

Duff, A., Boyle, E., Dunleavy, K., & Ferguson, J. (2004). The relationship between personality, approach to learning and academic performance. *Personality and Individual Differences*, *36*(8), 1907–1920.

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learning and Instruction*, *22*(4), 271–280.

Frey, R. F., Cahill, M. J., & McDaniel, M. A. (2017). Students' concept-building approaches: A novel predictor of success in chemistry courses. *Journal of Chemical Education*, *94*(9), 1185–1194.

Gentner, D., Loewenstein, J., & Thompson, L. (2003). Learning and transfer: A general role for analogical encoding. *Journal of Educational Psychology*, *95*(2), 393–408.

Gick, M. L., & Holyoak. K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, *15*, 1–38.

Gijbels, D., Van de Watering, G., Dochy, F., & Van den Bossche, P. (2005). The relationship between students' approaches to learning and the assessment of learning outcomes. *European Journal of Psychology of Education*, *20*(4), 327–341.

Groves, M. (2005). Problem-based learning and learning approach: Is there a relationship? *Advances in Health Sciences Education*, *10*(4), 315–326.

Holyoak, K. J., & Koh, K. (1987). Surface and structural similarity in analogical transfer. *Memory & Cognition*, *15*(4), 332–340.

Howie, P., & Bagnall, R. (2013). A critique of the deep and surface approaches to learning model. *Teaching in Higher Education*, *18*(4), 389–400.

Kek, M. Y. C. A., & Huijser, H. (2011). The power of problem-based learning in developing critical thinking skills: preparing students for tomorrow's digital futures in today's classrooms. *Higher Education Research & Development*, *30*(3), 329–341.

Kulasegaram, K. M., Chaudhary, Z., Woods, N., Dore, K., Neville, A., & Norman, G. (2017). Contexts, principles, and cognition: Principles for the transfer of basic science knowledge. *Medical Education*, *51*(2), 184–195.

Kulasegaram, K. M., Min, C., Ames, K., Howey, E., Neville, A., & Norman, G. (2012). The effect of principleual and contextual familiarity on transfer performance. *Advances in Health Sciences Education: Theory and Practice*, *17*(4), 489–499.

Kulasegaram, K. M., Min, C., Howey, E., Neville, A., Woods, N., Dore, K., & Norman, G. (2015). The mediating effect of context variation in mixed practice for transfer of basic science. *Advances in Health Sciences Education*, *20*(4), 953–968.

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into Practice*, *41*(4), 212–218.

Laird, T. F., Seifert, T. A., Pascarella, E. T., Mayhew, M. J., & Blaich, C. F. (2014). Deeply affecting first-year students' thinking: Deep approaches to learning and three dimensions of cognitive development. *The Journal of Higher Education*, *85*(3), 402–432.

Laird, T. F. N., Shoup, R., Kuh, G. D., & Schwarz, M. J. (2008). The effects of discipline on deep approaches to student learning and college outcomes. *Research in Higher Education*, *49*(6), 469–494.

McDaniel, M. A., Cahill, M. J., Frey, R. F., Rauch, M., Doele, J., Ruvolo, D., & Daschbach, M. M. (2018). Individual differences in learning exemplars versus abstracting rules: Associations with exam performance in college science. *Journal of Applied Research in Memory and Cognition*, *7*(2), 241–251.

McDaniel, M. A., Cahill, M. J., Robbins, M., & Wiener, C. (2014). Individual differences in learning and transfer: Stable tendencies for learning exemplars versus abstracting rules. *Journal of Experimental Psychology: General*, *143*(2), 668–693.

Nijhuis, J. F., Segers, M. S., & Gijselaers, W. H. (2005). Influence of redesigning a learning environment on student perceptions and learning strategies. *Learning Environments Research*, *8*(1), 67–93.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231–259.

Peirce, J. W. (2007). PsychoPy: Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1–2), 8–13.

Phan, H. P. (2011). Interrelations between self-efficacy and learning approaches: A developmental approach. *Educational Psychology*, *31*(2), 225–246.

Reeves, L., & Weisberg, R. W. (1994). The role of content and abstract information in analogical transfer. *Psychological Bulletin*, *115*(3), 381–400.

Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin*, *138*(2), 353–387.

Ross, B. H. (1987). This is like that: The use of earlier problems and the separation of similarity effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(4), 629–639.

Snelgrove, S., & Slater, J. (2003). Approaches to learning: Psychometric testing of a study process questionnaire. *Journal of Advanced Nursing*, *43*(5), 496–505.

Struyven, K., Dochy, F., Janssens, S., & Gielen, S. (2006). On the dynamics of students' approaches to learning: The effects of the teaching/learning environment. *Learning and Instruction*, *16*(4), 279–294.

Veenman, M. V., Prins, F. J., & Verheij, J. (2003). Learning styles: Self-reports versus thinking-aloud measures. *British Journal of Educational Psychology*, *73*(3), 357–372.

Watkins, D. (2001). Correlates of approaches to learning: A cross-cultural meta-analysis. In R. Sternberg & L. Zhang (Eds.)., *Perspective on thinking, learning, and cognitive styles* (pp. 165–195). New Jersey: Erlbaum.

Wilding, J., & Andrews, B. (2006). Life goals, approaches to study and performance in an undergraduate cohort. *British Journal of Educational Psychology*, *76*(1), 171–182.

Wilson, K., & Fowler, J. (2005). Assessing the impact of learning environments on students' approaches to learning: Comparing conventional and action learning designs. *Assessment & Evaluation in Higher Education*, *30*(1), 87–101.

Zeegers P. (2001) Student learning in science: A longitudinal study. *British Journal of Educational Psychology*, 71, 115–132.

Zhang, L. F. (2003). Does the big five predict learning approaches? Personality and Individual Differences, 34(8), 1431–1446.

Chapter 4: Do deep learning approaches predict transfer of knowledge? Operationally defining transfer outcomes in practice via student perceptions

The null correlation reported in Chapter 3 (Study 2) was obtained when we operationally defined transfer based on the perceptions of the instructor. Examining the robustness of these findings, we next sought to assess the predictive validity of the SPQ by operationally defining our performances measures based on the perceptions of students. To that end we administered the SPQ to students in two university courses: a lower-year health science course focusing on introductory epidemiology concepts, and an upper-year psychology course focusing on the diagnosis of mental illness in adolescents. After the final examinations of these courses, we invited a subset of students to rate each exam question on three constructs often associated with transfer: application of knowledge, contextual novelty, and the integration of multiple ideas or concepts. Performing a median split on these ratings, we then selectively examined performance on questions that students rated highly for each construct. In short, this study examined the predictive validity of the SPQ by operationally defining transfer through the lens of students.

Do deep learning approaches predict transfer of knowledge?

Operationally defining transfer outcomes in practice via student perceptions

Andrew B. LoGiudice[1], Geoffrey R. Norman[2], Saba Manzoor[3], & Sandra Monteiro[2]

[1]McMaster University, Department of Psychology, Neuroscience & Behaviour

[2]McMaster University, Health Research Methods, Evidence and Impact

[3]McMaster University, Faculty of Health Sciences

Correspondence to be addressed to Andrew B. LoGiudice, Department of Psychology, Neuroscience & Behaviour, McMaster University,1280 Main Street West, Hamilton, ON L8S-4K1. Phone: 289-339-3066. Email: logiudab@mcmaster.ca

Abstract

Students are often encouraged to focus on 'deep learning' by extracting general principles rather than memorizing details. As a testament to how entrenched this idea has become, inventories like the Study Process Questionnaire (SPQ) are now routinely used to quantify how much university students adopt so-called deep learning approaches during their courses. However, the SPQ's validity has been challenged due to vague descriptions of the outcome measures that serve as evidence of deep learning. Here we argue that such outcome measures should focus on the transfer of learning. To that end, we had university students complete the SPQ during one of their courses, write their final exam, then rate how much they perceived each final exam question to align with three constructs related to transfer: application of knowledge, contextual novelty, and integration of concepts. We then selectively examined performance on exam questions for which students reliably gave higher ratings—ostensibly reflecting more precise measures of transfer. Across two studies we found that SPQ deep approach scores did not predict or else very weakly predicted these transfer-oriented outcomes. These data further challenge the predictive validity of the SPQ, suggesting that educators and researchers should be cautious about using this tool to characterize student learning.

Keywords: Deep Learning; Approaches to Learning; Study Process Questionnaire; Transfer of Learning; Student Perceptions

Introduction

> [Understanding is] the interconnection of lots of disparate things . . . the way it all hangs together, the feeling that you understand how the whole thing is connected up . . . as though one's mind has finally 'locked in' to the pattern.
>
> —Entwhistle & Entwhistle (1992, p. 9)

Deep Learning

The idea of *deep learning* pervades discussions about higher education. Though it is hard to pinpoint a universally accepted definition, deep learning is usually described in terms of understanding or meaningful learning (e.g., Biggs, 1987a; Kember et al., 1997; Laird et al., 2008), higher-order learning (e.g., Agarwal, 2019; Bloom, 1956; Krathwohl, 2002), critical thinking (e.g., Kek & Huijser, 2011; Laird et al., 2014; Phan, 2011), or the application of general principles to new situations (e.g., Chernobilsky et al., 2004; Halpern & Hakel, 2003). Note how all these descriptions imply that deep learning reflects the generalization of abstract principles rather than mere memorization of details. As a testament to how entrenched this deep learning terminology has become, many researchers now claim that deep learning can be quantified in practice by asking students to report their study approaches via self-report inventories (for review, see Baeten et al., 2010), ultimately in hopes of evaluating the efficacy educational interventions (for review, see Kember et al., 1997).

The SPQ

The Study Process Questionnaire (SPQ) is perhaps the most popular inventory used to quantify deep learning (Biggs, 1987b; Biggs et al., 2001). Built upon theory from the student approaches to learning framework (e.g., Biggs, 1987a; Entwistle et al, 1979; Marton & Säljö, 1976), the SPQ prompts a student with Likert items about their study attitudes or habits within a particular course, in turn producing a 'deep approach' score that is said to quantify the student's tendency to adopt deep learning throughout said course. Definitions of deep learning vary by author, but apt examples include processing content "at a high level of generality, such as main ideas, themes, and principles rather than as conceptually unsupported specifics" (Biggs, 1993, p. 7) or "understanding core concepts [to see] relationships among them or figuring out how to apply information in new ways" (Laird et al., 2014, p. 405). Put simply, the main thrust of the SPQ is that its deep approach measure reflects a student's tendency to extract general principles

within a course, and so this measure ought to predict performance on tasks that require this so-called 'deep' form of learning.

Despite the SPQ's widespread use, however, there is arguably little evidence of its predictive validity. Findings are mixed, with some studies showing positive correlations between SPQ deep approach scores and academic achievement (e.g., McManus et al., 1998; Snelgrove & Slater, 2003; Zeegers, 2001) and many others showing no correlations (e.g., Gijbels et al., 2005; Groves, 2005; Wilding & Andrews, 2006). Shedding light on this issue, a large meta-analysis by Richardson et al. (2012) found that deep approach scores correlated only very weakly with overall academic achievement in university students ($r^+ = .14$, $p = .03$), accounting for less than 3% of variance. But these mixed findings and the weak correlation found in the meta-analysis might be an artefact of SPQ studies largely relying on aggregate outcome measures (e.g., final exam performance, cumulative GPA) that are highly heterogenous, and hence may not precisely reflect whether students extracted general principles per se (Watkins, 2001, as cited in Choy et al., 2012). There is also much ambiguity in what specific learning outcomes serve as evidence of deep learning (Dinsmore & Alexander, 2012; Howie & Bagnall, 2013). Therefore, the real question is whether the SPQ's deep approach scores positively correlate with outcome measures that more precisely index how well students extracted general principles from the course content.

Transfer-Oriented Outcomes

We argue that the SPQ's validity depends on how well it predicts *transfer of learning*—i.e., the process by which learners solve novel problems by generalizing from their past learning experiences (for review, see Day & Goldstone, 2012). Deep learning and transfer share many obvious similarities: both are considered desirable yet difficult (e.g., Gick & Holyoak, 1983; Norman, 2009; Sandberg & Barnard, 1997); both entail learning of general principles and subsequent application to new problems (e.g., Chernobilsky et al., 2004; Gentner & Smith, 2012; Needham & Begg, 1991); both entail the learner understanding how these general principles extend to novel contexts (e.g., Barnett & Ceci, 2002; Biggs, 1999; Catrambone & Holyoak, 1989); and both entail the integration of multiple related ideas or concepts (e.g., Agarwal, 2019; Biggs, 1999; Butler, 2010). We emphasize that this comparison is not new; indeed, many others have noted the strong conceptual link between deep learning and transfer or learning (e.g., Hattie & Donoghue, 2016; Laird et al., 2008). It therefore seems surprising that more work has not tried to validate the SPQ using outcome measures that more explicitly tap transfer.

Present Studies

To that end, here we tested the SPQ's validity by operationally defining our outcome measures based on student perceptions of constructs commonly associated with transfer. The SPQ relies on the assumption that students hold intuitive notions regarding varied approaches to learning in relation to deep learning. Therefore, it was necessary to test that assumption by asking students to estimate how much of their exam questions required 'application' of course content. We additionally assumed that students could estimate how 'novel' each exam question felt relative to prior course content or assessments, given that they are directly engaged with the course content. (e.g., through global feelings of familiarity; Whittlesea et al., 1990; Yonelinas, 2001), and how much each exam question forced them to 'integrate' multiple ideas or concepts (e.g., through perceived effort or difficulty; DeLeeuw & Mayer, 2008; Paas et al., 2003). We reasoned that performance on these refined subsets of exam questions might serve as more precise measures of transfer with which to assess the SPQ's validity.

We were guided by one simple prediction: If the SPQ captures a student's likelihood of extracting general principles, then its deep approach scores ought to correlate positively with student final exam performance—particularly for subsets of exam questions that students perceived to more closely align with the constructs of application, novelty, or integration.

Study 1

Method

Participants

We selected a course with 366 students from a second-year university course that focused on introductory concepts in epidemiology. Subtopics included study designs, standard measures of health, indices of association, and specialized research topics relevant to the health sciences (e.g., evidence-based medicine, systematic reviews, control of infectious diseases). The study was approved by the [Blinded for Review] Research Ethics Board (Protocol #0591).

Two hundred and seventy-eight students (72% female) consented to participate by completing the survey; a 76% response rate. All these students completed the SPQ during the course and gave us permission to access their final exam grades. Most students (78%) were in the second year of their program, some (21%) in their third year, and the remainder (1%) in their

first year. As noted in the paragraph above, twelve of these students volunteered to return after the final exam to rate the exam questions.

Materials

Study Process Questionnaire. We adopted the most popular variant of the SPQ, the R-SPQ-2F (Biggs et al., 2001). This questionnaire consists of 20 Likert items, 10 corresponding to the deep approach construct, and 10 to the surface approach construct (this latter construct relates to rote memorization and other superficial study strategies). Note that we were only interested in the deep approach construct here because it more closely represents the goal of educators. Each of the 10 items corresponding to the deep approach construct requires a response from 1–5 so that each student receives a deep approach score from 0–50, with higher scores meant to reflect deeper learning approaches.

Items for Rating Exam Questions. Based on the three target constructs, we created three Likert items, each prompting students to rate how strongly one of the three constructs aligned with each of the exam questions (Appendix A). The first item adopted typical deep learning terminology, asking students to rate how much each question required them to apply course content "at a high level of generality, such as main ideas, themes, and principles" (Biggs, 1993, p. 7); the second item asked students to rate how novel each question felt relative to prior course content; and the third item asked students to rate how much they thought each question required multiple ideas to be integrated. Each item required a response from 1–4 such that higher ratings reflect greater alignment with the targeted construct.

Procedure

A month before the course's final exam, we announced to students that they could complete an online survey about their study approaches to earn a small bonus grade. Students were permitted to complete the survey anytime between the announcement and the final exam. The survey asked them to report their sex, what year of studies they were in, approximately how many class sessions they attended, and to respond to all 20 items from the R-SPQ-2F with respect to their study approaches for the course. The survey also asked them whether they consented to letting us access their final exam grades.

After these students wrote the final exam—a total of 81 multiple choice questions—we invited a randomly selected subset of 12 students to share their perceptions of the exam

questions. Seven returned one day after the exam and five returned twelve days after the exam. These students first completed the SPQ again; this allowed us to examine its test–retest reliability with a mean interval of 35.0 days (SD = 11.9) between tests. They then received a survey that asked them to rate each question on the final exam using the three Likert items (Appendix A) measuring application, novelty, and integration. These student ratings of exam questions allowed us to examine how much students agreed in their perceptions of the three target constructs. Most critically, these ratings also allowed us to selectively examine student performance on subsets of exam questions that students rated highly for each construct. A schematic of the study design is shown in Figure 1.

< Figure 1 here >

Results

We organized our results into three sections: data quality, predicted relations, and exploratory relations. Correlations are reported as Pearson's *r* and contrasts between means are reported as *t*-tests, with both using two-tailed tests and a significance criterion of $\alpha = .05$.

Data Quality

Reliability of SPQ Deep Approach Scores. Test–retest reliability for the SPQ's deep approach scores was calculated using a two-way random effects model. Reliability was good (ICC = .85), indicating that these scores could distinguish between students based on their self-reported approaches to learning. Because the SPQ data achieved this level of reliability, they were deemed suitable for detecting the predicted positive correlation with exam performance.

Reliability of Aggregate Exam Scores. Internal consistency of the final examination was calculated using a two-way mixed effects model. Reliability was good (ICC = .83), indicating that this aggregate outcome measure was a reliable estimate of students' learning of the course content. Based on this acceptable level of reliability, the exam scores were deemed suitable for detecting the predicted positive correlation with deep approach scores.

Reliability of Application, Novelty, and Integration Ratings. Inter-rater reliability of student ratings of application, novelty, and integration was calculated using a two-way random effects model based on mean ratings of the 12 randomly selected students (k = 12). Reliability was good for ratings of the application construct (ICC = .84), moderate for ratings of the novelty construct (ICC =.64), and good for ratings of the integration construct (ICC =.84). Thus, ratings

were relatively consistent between students and effectively discriminated between exam questions, justifying use of the three constructs to classify exam questions for further analysis.

Reliability of Transfer-Oriented Exam Scores. We then used these student ratings of exam questions to create outcome measures that more precisely reflect transfer, at least as perceived by students. Three median splits were performed on all the exam questions, once for each of the three ratings items, with replacement. Exam questions with a mean rating equal to the median rating were excluded. Ratings derived from the first item yielded 40 exam questions that we labelled 'high application' and 35 that we labelled 'low application'; ratings from the second item yielded 35 exam questions that we labelled 'high novelty' and 40 that we labelled 'low novelty'; and ratings from the third item yielded 36 exam questions that we labelled 'high integration' and 38 that we labelled 'low integration'. We were primarily interested in whether SPQ deep approach scores could specifically predict performance on these highly rated subsets of questions because students reliably associated them with the three target constructs, and so the remaining questions that were rated lowly were not included in further analysis.

To ensure that these highly rated subsets of questions served as reliable outcome measures, we next calculated internal consistency using a two-way mixed effects model. Inter-item reliability was good for exam questions rated highly on application (k = 40, ICC = .78), good for exam questions rated highly on novelty (k = 35, ICC = .75), and moderate for exam questions rated highly on integration (k = 36, ICC = .70). All three outcome measures were thus suitable for detecting the predicted positive correlations with SPQ deep approach scores.

Predicted Relations

Predictive Validity for Aggregate Exam Scores. Consistent with prior work, we computed correlations between SPQ deep approach scores and aggregate exam scores. The correlation was nearly zero (r = -.02, p = .741). This brings us to our more precise research question: Do SPQ deep approach scores more specifically predict performance for the subsets of exam questions that students perceived as reflecting the constructs of application, novelty, or integration?

Predictive Validity for Transfer-Oriented Exam Scores. Figure 2 shows correlations between SPQ deep approach scores and the three transfer-oriented outcome measures. There was no significant correlation for exam questions rated highly on application (r = -.03, p = .618),

novelty ($r = -.01$, $p = .910$), or integration ($r = .03$, $p = .571$). Thus, despite our efforts to create outcome measures that more precisely reflect transfer, we found no evidence for the SPQ's predictive validity.

< Figure 2 here >

Exploratory Relations

Correlations Between Student Ratings. We next explored how much student ratings of the three target constructs (application, novelty, integration) overlapped. Figure 3 shows correlations between ratings of each construct. Perceptions of application and novelty were weakly positively correlated ($r = .26$, $p = .017$), perceptions of application and integration were strongly positively correlated ($r = .76$, $p < .001$), and perceptions of integration and novelty were moderately positively correlated ($r = .55$, $p < .001$). Thus, student perceptions of application and integration were closely linked, whereas perceptions of both these constructs were relatively distinct from perceptions of novelty.

< Figure 3 here >

Exam Performance as a Function of Student Ratings. Lastly, we explored whether performance on exam questions varied as a function of classification into the 'high' or 'low' category for each target construct. We did this by performing three paired-sample t-tests on exam performance, with 'high' and 'low' labels for each construct derived from the three median splits. Figure 4 reveals a consistent trend: higher ratings of application ($p < .001$, $d = .46$), novelty ($p < .001$, $d = .99$), and integration ($p < .001$, $d = 1.6$) were all associated with poorer exam performance. At face value, this pattern of results is consistent with the idea that transfer is difficult, reinforcing the view that these student ratings reflected meaningful constructs associated with transfer.

< Figure 4 here >

Discussion

The SPQ's deep approach scores did not predict student performance on exam questions that students collectively perceived to resemble transfer. Despite the relatively high reliability observed for SPQ deep approach scores, for student ratings of the three constructs, and for our transfer-oriented outcome measures, there was no evidence of predictive validity.

One possibility is that either the course content or the exam questions did not effectively capture what researchers view as transfer. After all, this was an introductory course, and so it may have focused on memorization of new and unfamiliar terms rather than a richer understanding of general principles. The final exam was also comprised exclusively of multiple-choice questions, which some argue tend to tap into rote memorization or familiarity with key terms rather than an understanding of general principles (e.g., Ferland et al., 1987; Frederiksen, 1984; Stanger-Hall, 2012; cf. Little et al., 2012). Therefore, before drawing conclusions, we wished to examine the generalizability of our findings by conducting the same study again, but this time using an upper-year course with a final exam comprised mainly of written long-answer questions.

Study 2

Method

Participants

We selected a third-year university course of 194 students that focused on mental health in young children and adolescents. Subtopics included symptomology, diagnostic criteria, and treatment of mental health disorders. One hundred and forty students (86% female) consented to participate in the study; a 73% response rate. All these students completed the SPQ during the course and gave us permission to access their final exam grades. Most (54%) were in the third year of their program, some (35%) in their fourth year, and the remainder (11%) in their fifth year or higher. Seven students returned one day after the final exam and five students returned twelve days after the final exam with a mean interval of 51.3 days (SD = 14.1) between administrations of the SPQ. The study was approved by the [Blinded for Review] Research Ethics Board (Protocol #0591).

Procedure

In contrast to Study 1, the 52-question final exam for this course mainly consisted of long-answer type questions for which students had to provide written responses with multiple steps (e.g., analyzing novel patient cases with the explicit instructor-stated goal of integrating multiple course concepts). After the final exam, 12 students returned to rate the exam questions in the same manner as Study 1.

Results

Data Quality

Reliability of SPQ Deep Approach Scores. Test–retest reliability for the SPQ's deep approach scores was calculated using a two-way random effects model. Reliability was good (ICC = .81), indicating that these scores were suitable for detecting the predicted correlation between SPQ scores and exam performance.

Reliability of Aggregate Exam Scores. Internal consistency of the final examination was calculated using a two-way mixed effects model. Reliability was good (ICC = .86), indicating that this aggregate outcome measure was suitable for detecting a correlation with SPQ scores.

Reliability of Application, Novelty, and Integration Ratings. Inter-rater reliability of student ratings of application, novelty, and integration was calculated using a two-way random effects model based on mean ratings of the 12 randomly selected students (k = 12). Inter-rater reliability for the mean of 12 ratings was good for the first item related to application of knowledge (ICC = .84), moderate for the second item related to novelty (ICC =.69), and good for the third item related to integration of multiple concepts (ICC =.87). As in Study 1, ratings pertaining to the three constructs were relatively consistent between students and effectively discriminated between exam questions, justifying the use of these ratings for further analysis.

Reliability of Transfer-Oriented Exam Scores. Three median splits were again performed on all the exam questions in the same manner as Study 1. Ratings from the first item yielded 25 exam questions that we labelled 'high application' and 24 that we labelled 'low application'; ratings from the second item yielded 22 exam questions that we labelled 'high novelty' and 25 that we labelled 'low novelty'; and ratings from the third item yielded 25 exam questions that we labelled 'high integration' and 25 that we labelled 'low integration'.

To ensure the three highly rated subsets of questions produced reliable outcome measures, we next calculated internal consistency using a two-way mixed effects model. Inter-item reliability was good for exam questions rated highly on application (k = 25, ICC = .77), good for exam questions rated highly on novelty (k = 22, ICC = .78), and good for exam questions rated highly on integration (k = 25, ICC = .78). All three outcome measures were

therefore suitable for detecting the predicted positive correlations with SPQ deep approach scores.

Predicted Relations

Predictive Validity for Aggregate Exam Scores. Contrary to Study 1, here the correlation between SPQ scores and aggregate exam performance was positive albeit small, and trending toward statistical significance ($r = .17$, $p = .051$). This positive but weak correlation raised the interesting possibility that we might detect stronger correlations using our refined transfer-oriented outcome measures.

Predictive Validity for Transfer-Oriented Exam Scores. Correlations between SPQ deep approach scores and each of the three outcome measures are depicted in Figure 5. There were weak but marginally significant positive correlation between SPQ deep approach scores and performance on exam questions rated highly on application ($r = .16$, $p = .054$). There were also weak and statistically significant correlations with SPQ deep approach scores for performance on exam questions rated highly on novelty ($r = .17$, $p = .046$) and integration ($r = 18$, $p = .032$). We therefore found some weak evidence to suggest that the SPQ was able to predict transfer-oriented learning outcomes, though the correlations were very small.

< Figure 5 here >

Exploratory Relations

Correlations Between Student Ratings. As shown in Figure 6, ratings of application were modestly correlated with ratings of novelty ($r = .43$, $p = .001$), ratings of application were highly correlated with ratings of integration ($r = .87$, $p < .001$), and there was also a modest correlation between integration and novelty ($r = .49$, $p < .001$). Once again, student perceptions of application and integration were closely linked, whereas perceptions of both these constructs were relatively distinct from perceptions of novelty.

< Figure 6 here >

Exam Performance as a Function of Students Ratings. Lastly, we explored whether performance on exam questions varied as a function of classification into the ‘high’ or ‘low’ category for each target construct. We did this by performing three paired-sample t-tests on exam performance, with ‘high’ and ‘low’ labels for each construct derived from the three median splits. Figure 7 reveals the same trend observed in Study 1: higher ratings of application ($p <$

.001, d = 1.6), novelty (p < .001, d = 1.6), and integration (p < .001, d = .76) were all associated with poorer exam performance.

< Figure 7 here >

General Discussion

A similar pattern of results was found across two different courses: (a) SPQ deep approach scores were reliable; (b) student ratings of exam questions in terms the three transfer-oriented constructs were reliable; (c) using these student ratings to select subsets of exam questions produced three reliable outcome measures that were perceived to resemble transfer; and (d) SPQ deep approach scores did not predict, or else marginally predicted, these transfer-oriented outcome measures. The strongest evidence for the predictive validity of the SPQ came from weak but significant correlations in Study 2, yet even here its deep approach scores only accounted for approximately 3% of variance in our transfer-oriented outcome measures.

A novel contribution of these studies was the method of refining outcome measures based on student perceptions of three transfer-oriented constructs. Using 12 students per course, we obtained reliable estimates of how much students perceived each exam question to resemble each of the three constructs. Because these findings were also replicated across two different courses, we propose that future studies might benefit from adopting a similar student-centered approach for operationally defined measures of transfer in authentic educational settings.

Our exploratory analyses also revealed some serendipitous findings that might elucidate relations between the three transfer-oriented constructs. First, despite the somewhat vague terminology used to describe deep learning, students generally agreed in their perceptions of which exam questions required them to apply knowledge at a high level of generality, thus providing empirical support for the construct of deep learning. However, while deep learning can be associated with specific question characteristics, the link to actual study strategies, which is an implicit assumption of the SPQ, seems more tenuous. Second, it was interesting that perceptions of application closely aligned with perceptions of integration. This result suggests that deep learning (as typically described by researchers) might closely align with the number of ideas or concepts that must be integrated. Third, perceptions of novelty were reliable, but relatively distinct from perceptions of both application and integration. This result suggests novelty represents a separate but meaningful construct related to transfer. These three findings speak to

the multifaceted nature of the deep learning and transfer constructs. And in that vein, we suspect future studies might benefit from adopting multiple outcome measures that reflect each of the three constructs investigated here, perhaps elucidating how the constructs are interrelated.

Limitations

A limitation of this work is that many inventories exist to quantify deep learning, and yet here we only used the R-SPQ-2F. This leaves open the possibility that other similar learning inventories might be more effective at predicting similar transfer-oriented outcomes. Nonetheless, the R-SPQ-2F is arguably the most popular of these inventories, and so the present data speak to at least a large body of research in higher education investigating deep learning.

Another limitation is that we relied solely on student perceptions to create our refined transfer-oriented outcome measures. This concern might be allayed by considering that student ratings of the three constructs were reliable, suggesting these ratings represented meaningful constructs. Our exploratory analyses also revealed that performance was substantially worse for exam questions that students rated highly for each construct, consistent with what one might expect if these subsets of exam questions did effectively tap into transfer.

Conclusion

We suspect that the popularity of the SPQ stems from its intuitive appeal and ease of use. However, such tools to quantify deep learning are only productive insofar as the specific student learning outcomes that evince deep learning are well-defined (see Howie & Bagnall, 2013, for a similar argument). Here we made a small step in this direction by harnessing student perceptions of transfer, creating outcome measures to more precisely measure whether students extracted general principles from the course content. And yet, even with these refined transfer-oriented outcome measures, we failed to find substantial evidence for the SPQ's predictive validity. We conclude that deep learning is a fascinating construct, but that the available data—including the data reported here—do not convincingly show that the SPQ captures a student's tendency to extract generalizable principles. Unless more compelling evidence is produced in future studies, we caution researchers against using this tool to characterize student learning.

Acknowledgements

This research was supported in part by a SSHRC Canada Graduate Scholarship (Doctoral) awarded to the first author. We thank two instructors at McMaster University, Russel de Souza and Laura Jin, for allowing us to conduct these studies within their courses.

References

Agarwal, P. K. (2019). Retrieval practice & Bloom's taxonomy: Do students need fact knowledge before higher order learning? *Journal of Educational Psychology, 111*(2), 189–209.

Baeten, M., Kyndt, E., Struyven, K., & Dochy, F. (2010). Using student-centred learning environments to stimulate deep approaches to learning: Factors encouraging or discouraging their effectiveness. *Educational Research Review, 5*(3), 243–260.

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin, 128*(4), 612–637.

Biggs, J. (1993). What do inventories of students' learning processes really measure? A theoretical review and clarification. *British Journal of Educational Psychology, 63*(1), 3–19.

Biggs, J. (1999). What the student does: Teaching for enhanced learning. *Higher Education Research & Development, 18*(1), 57–75.

Biggs, J. B. (1987a). *Student Approaches to Learning and Studying.* Hawthorn, Australia: Australian Council for Educational Research Ltd.

Biggs, J. B. (1987b). *Study Process Questionnaire Manual. Student Approaches to Learning and Studying*. Hawthorn, Australia: Australian Council for Educational Research Ltd.

Biggs, J., Kember, D., & Leung, D. Y. (2001). The revised two-factor study process questionnaire: R-SPQ-2F. *British Journal of Educational Psychology, 71*(1), 133–149.

Bloom, B.S. (Ed.), Engelhart, M.D., Furst, E.J., Hill, W.H., & Krathwohl, D.R. (1956). *Taxonomy of educational objectives: The classification of educational goals*. Handbook 1: Cognitive domain. New York: David McKay.

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(5), 1118–1133.

Catrambone, R., & Holyoak, K. J. (1989). Overcoming contextual limitations on problem-solving transfer. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*(6), 1147–1156.

Chernobilsky, E., DaCosta, M. C., & Hmelo-Silver, C. E. (2004). Learning to talk the educational psychology talk through a problem-based course. *Instructional Science, 32*(4), 319–356.

Choy, J. L. F., O'Grady, G., & Rotgans, J. I. (2012). Is the Study Process Questionnaire (SPQ) a good predictor of academic achievement? Examining the mediating role of achievement-related classroom behaviours. *Instructional Science, 40*(1), 159–172.

Day, S. B., & Goldstone, R. L. (2012). The import of knowledge export: Connecting findings and theories of transfer of learning. *Educational Psychologist, 47*(3), 153–176.

DeLeeuw, K. E., & Mayer, R. E. (2008). A comparison of three measures of cognitive load: Evidence for separable measures of intrinsic, extraneous, and germane load. *Journal of Educational Psychology, 100*(1), 223–234.

Dinsmore, D. L., & Alexander, P. A. (2012). A critical discussion of deep and surface processing: What it means, how it is measured, the role of context, and model specification. *Educational Psychology Review, 24*(4), 499–567.

Entwistle, A., & Entwistle, N. (1992). Experiences of understanding in revising for degree examinations. *Learning and Instruction*, *2*(1), 1–22.

Entwistle, N., & McCune, V. (2004). The conceptual bases of study strategy inventories. *Educational Psychology Review*, *16*(4), 325–345.

Entwistle, N., Hanley, M., & Hounsell, D. (1979). Identifying distinctive approaches to studying. *Higher Education*, *8*(4), 365–380.

Ferland, J. J., Dorval, J., & Levasseur, L. (1987). Measuring higher cognitive levels by multiple choice questions: A myth? *Medical Education*, *21*(2), 109–113.

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist*, *39*, 193–202.

Gentner, D. & Smith, L. (2012). Analogical reasoning. In V. S. Ramachandran (Ed.), *Encyclopedia of Human Behavior* (2nd Ed., pp. 130–136). Oxford, UK: Elsevier.

Gick, M. L., & Holyoak. K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, *15*, L–38.

Gijbels, D., Van de Watering, G., Dochy, F., & Van den Bossche, P. (2005). The relationship between students' approaches to learning and the assessment of learning outcomes. *European Journal of Psychology of Education*, *20*(4), 327–341.

Groves, M. (2005). Problem-based learning and learning approach: Is there a relationship? *Advances in Health Sciences Education*, *10*(4), 315–326.

Halpern, D. F., & Hakel, M. D. (2003). Applying the science of learning to the university and beyond: Teaching for long-term retention and transfer. *Change: The Magazine of Higher Learning*, *35*(4), 36–41.

Hattie, J. A., & Donoghue, G. M. (2016). Learning strategies: A synthesis and conceptual model. *NPJ Science of Learning*, *1*(1), 1–13.

Howie, P., & Bagnall, R. (2013). A critique of the deep and surface approaches to learning model. *Teaching in Higher Education*, *18*(4), 389–400.

Kek, M. Y. C. A., & Huijser, H. (2011). The power of problem-based learning in developing critical thinking skills: Preparing students for tomorrow's digital futures in today's classrooms. *Higher Education Research & Development*, *30*(3), 329–341.

Kember, D. (1997). Evaluating the effectiveness of educational innovations: Using the study process questionnaire to show that meaningful learning occurs. *Studies in Educational Evaluation*, *23*(2), 141–57.

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into Practice*, *41*(4), 212–218.

Laird, T. F. N., Seifert, T. A., Pascarella, E. T., Mayhew, M. J., & Blaich, C. F. (2014). Deeply affecting first-year students' thinking: Deep approaches to learning and three dimensions of cognitive development. *The Journal of Higher Education*, *85*(3), 402–432.

Laird, T. F. N., Shoup, R., Kuh, G. D., & Schwarz, M. J. (2008). The effects of discipline on deep approaches to student learning and college outcomes. *Research in Higher Education*, *49*(6), 469–494.

Little, J. L., Bjork, E. L., Bjork, R. A., & Angello, G. (2012). Multiple-choice tests exonerated, at least of some charges: Fostering test-induced learning and avoiding test-induced forgetting. *Psychological Science*, *23*(11), 1337–1344.

Marton, F., & Säljö, R. (1976). On qualitative differences in learning: I: Outcome and process. *British Journal of Educational Psychology*, *46*(1), 4–11.

McManus, I. C., Richards, P., Winder, B. C., & Sproston, K. A. (1998). Clinical experience, performance in final examinations, and learning style in medical students: Prospective study. *British Medical Journal*, *316*(7128), 345–350.

Needham, D. R., & Begg, I. M. (1991). Problem-oriented training promotes spontaneous analogical transfer: Memory-oriented training promotes memory for training. *Memory & Cognition*, *19*(6), 543–557.

Norman, G. (2009). Teaching basic science to optimize transfer. *Medical Teacher*, *31*(9), 807–811.

Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist*, *38*(1), 63–71.

Phan, H. P. (2011). Deep processing strategies and critical thinking: Developmental trajectories using latent growth analyses. *The Journal of Educational Research*, *104*(4), 283–294.

Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin*, *138*(2), 353–387.

Sandberg, J., & Barnard, Y. (1997). Deep learning is difficult. *Instructional Science*, *25*(1), 15–36.

Schwartz, D. L., Chase, C. C., Oppezzo, M. A., & Chin, D. B. (2011). Practicing versus inventing with contrasting cases: The effects of telling first on learning and transfer. *Journal of Educational Psychology*, *103*(4), 759–775.

Snelgrove, S., & Slater, J. (2003). Approaches to learning: Psychometric testing of a study process questionnaire. *Journal of Advanced Nursing*, *43*(5), 496–505.

Stanger-Hall, K. F. (2012). Multiple-choice exams: An obstacle for higher-level thinking in introductory science classes. *CBE: Life Sciences Education*, *11*(3), 294–306.

Watkins, D. (2001). Correlates of approaches to learning: A cross-cultural meta-analysis. In R. Sternberg & L. Zhang (Eds.)., *Perspective on thinking, learning, and cognitive styles* (pp. 165–195). New Jersey: Erlbaum.

Whittlesea, B. W., Jacoby, L. L., & Girard, K. (1990). Illusions of immediate memory: Evidence of an attributional basis for feelings of familiarity and perceptual quality. *Journal of Memory and Language*, *29*(6), 716–732.

Wilding, J., & Andrews, B. (2006). Life goals, approaches to study and performance in an undergraduate cohort. *British Journal of Educational Psychology*, *76*(1), 171–182.

Yonelinas, A. P. (2001). Components of episodic memory: the contribution of recollection and familiarity. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *356*(1413), 1363–1374.

Zeegers, P. (2001). Approaches to learning in science: A longitudinal study. *British Journal of Educational Psychology*, *71*(1), 115–132.

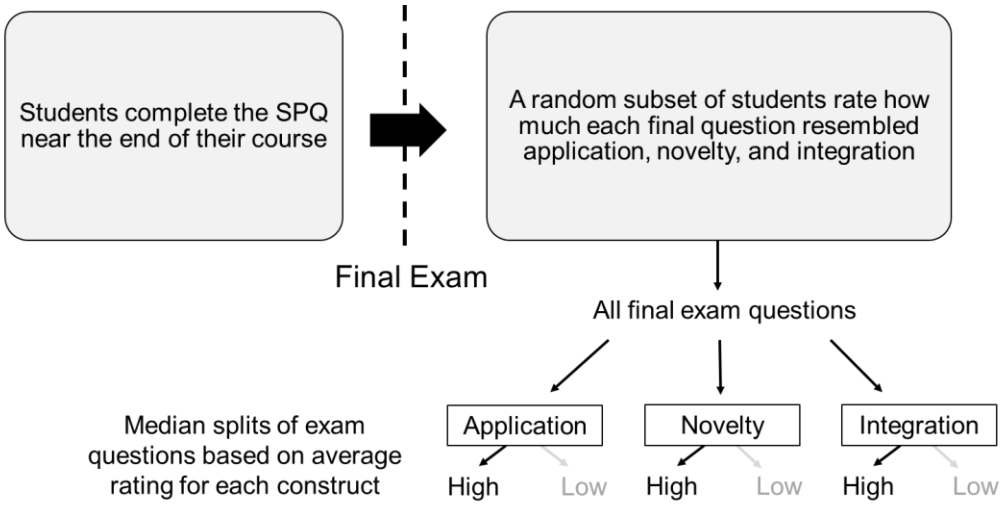


Figure 1. A schematic of the core design adopted in Study 1 and Study 2.

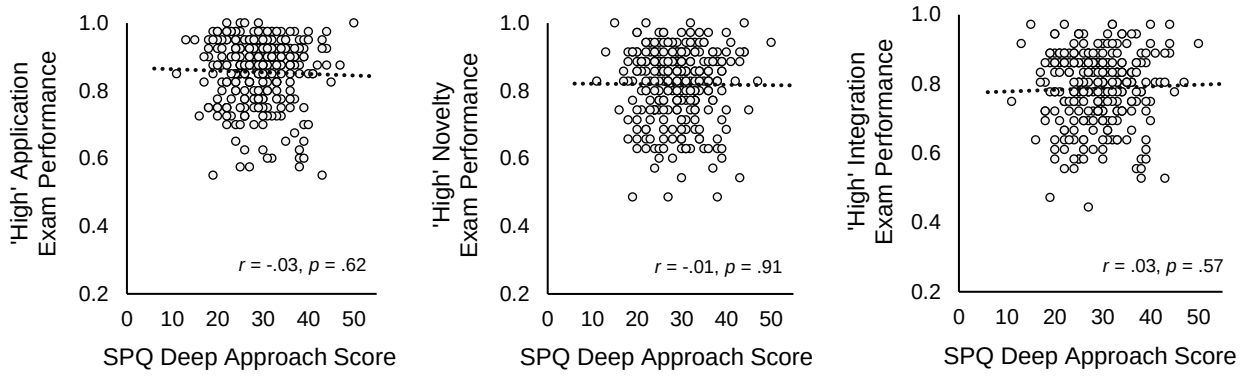


Figure 2. Correlations between SPQ deep approach scores and performance on exam questions that students rated highly in terms of application, novelty, and integration in Study 1 (N = 278).

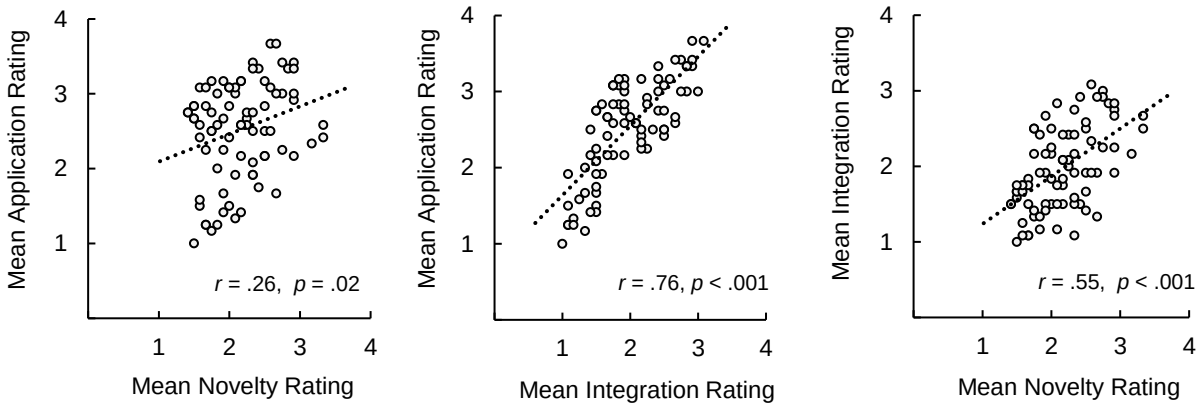


Figure 3. Pairwise correlations between mean student ratings of each construct (application, novelty, integration) for all 81 final exam questions in Study 1.

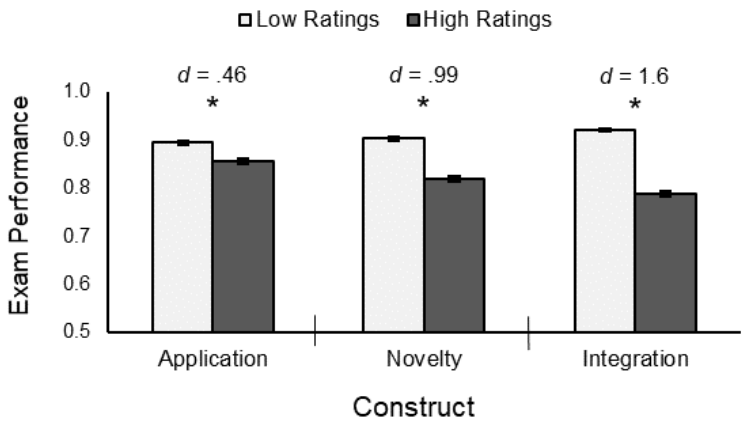


Figure 4. Performance on exam questions as a function of their classification as 'high' or 'low' for the three transfer-oriented constructs in Study 1. Error bars denote ± SEM.

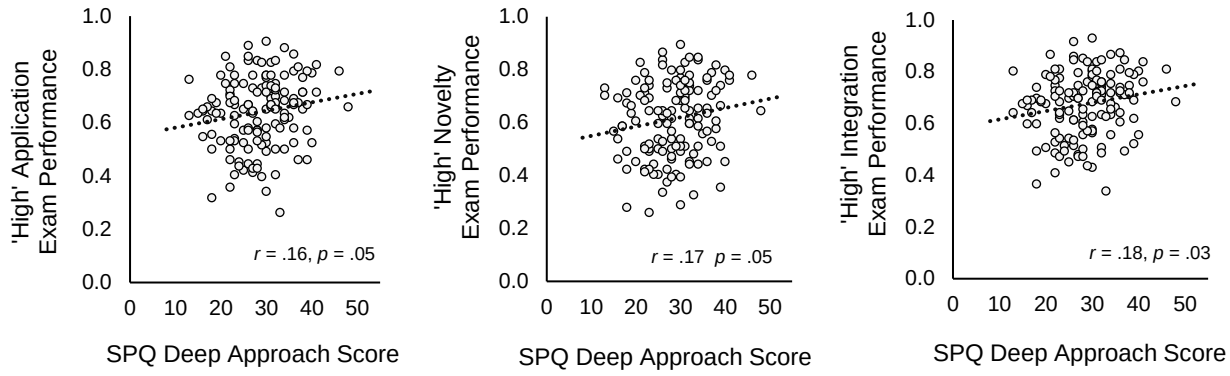


Figure 5. Correlations between the SPQ deep approach scores of students and for performance for subsets of final exam questions that they rated highly in terms of the three transfer-oriented constructs (application, novelty, integration) in Study 2 (N = 140).

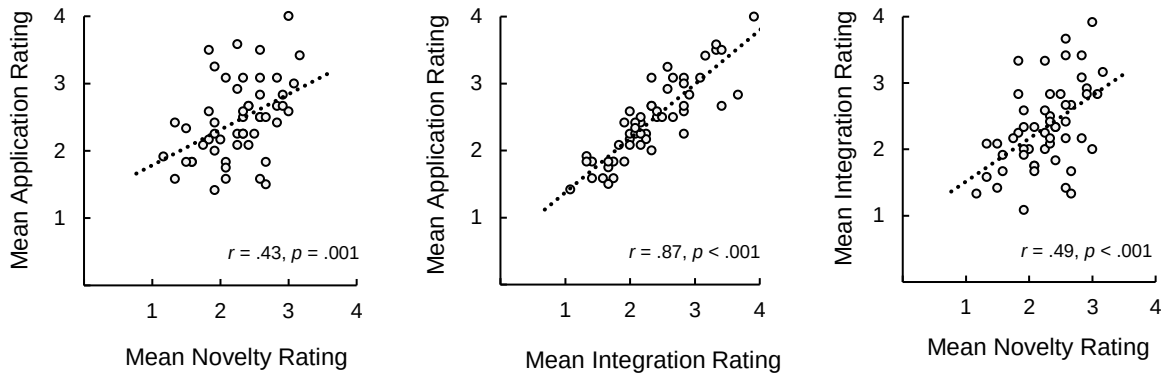


Figure 6. Pairwise correlations between student ratings of each construct (application, novelty, integration) for all 52 final exam questions in Study 2.

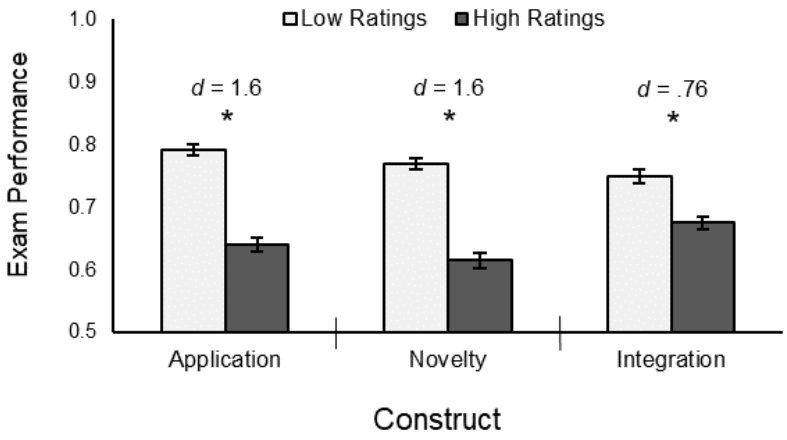


Figure 7. Performance on exam questions as a function of their classification as 'high' or 'low' for the three transfer-oriented constructs in Study 2. Error bars denote ± SEM.

Appendix A

Each letter below (A, B, C) corresponds to a column in the provided excel document. Enter a number from 1–4 in every cell.

Please use the full range of each scale;
and don't hesitate to use the far ends (1 and 4) as you deem appropriate!

A) How much did this question require you to **apply** course content at a high level of generality using main ideas, themes, or principles?

no application	1	2	3	4	**a lot of application**

B) How similar or **novel** did this specific question seem to other specific examples, practice questions, or test questions that you saw previously during the course?

similar	1	2	3	4	**novel**

C) How much did this question require you to **integrate** multiple ideas or concepts that you learned throughout the course?

no integration	1	2	3	4	**a lot of integration**

Chapter 5: General Discussion

The goal of this final chapter is to situate all the findings from this thesis into the broader literature. For the experimental approach I elaborate upon the findings from Chapter 2 in relation to the retrieval-based learning literature, outlining different mechanisms through which effortful retrieval attempts might promote learning and subsequent transfer. This focus on underlying mechanisms is critical; studies on retrieval practice have been criticized for lacking a strong theoretical framework, so there is an impetus toward exploring moderating variables and boundary conditions using learning tasks that go beyond verbatim recall (e.g., Butler et al., 2017; van Gog & Sweller, 2015; Rawson, 2015; Woolridge et al., 2014). It is my hope that the findings from Chapter 2 will help us better understand these moderating variables and boundary conditions.

Next, for the correlational approach I elaborate on the findings from Chapters 3 and 4 in relation to ongoing discussions about the SPQ and other similar learning inventories. Most of this section is dedicated to contrasting the SPQ with other information-processing accounts that have been developed to explain findings with simpler learning tasks. I also discuss the challenges of relying on self-report inventories to gauge student learning approaches. Most critically, I argue there is little to be gained using the SPQ or similar self-report inventories unless greater emphasis is placed on specific cognitive processes and their consequences.

The Experimental Approach

Evidence for Variability-Induced Retrieval

The experiments in Chapter 2 began with two explanations for why contextual variability can promote learning and subsequent transfer. The variability-induced encoding hypothesis suggests that variable contexts during training are beneficial because they encourage participants to encode the conserved deep structure across training instances rather than nonessential details. On the other hand, the variability-induced retrieval hypothesis suggests that variable contexts during training can be beneficial because they induce more effortful retrieval of similar prior training events stored in memory.

Studies of analogical transfer nearly always focus on the variability-induced encoding hypothesis, as discussed in Chapter 1. The standard design consists of two groups that differ in training: an experimental group that receives multiple examples shown simultaneously with

explicit instructions to compare them, and a control group that receives the same examples in sequence without any such comparison instructions. The experimental group usually outperforms control group on a subsequent transfer test, presumably because comparing examples facilitates the discernment of deep structure via structural alignment (for review, see Gentner & Smith, 2012). But note how this typical design overlooks the variability-induced retrieval hypothesis: the control condition may involve training instances that are spaced in time, but they are usually presented as to-be-studied examples rather than to-be-solved problems, minimizing the need to explicitly retrieve instances that were studied in the past.

The goal of Chapter 2 was therefore to test the benefit of variability-induced retrieval *above and beyond* any benefit of variability-induced encoding. To our surprise, however, we found no evidence for variability-induced encoding in the adjacent condition where one might expect participants to spontaneously engage in structural alignment. On the other hand, we did find evidence for the variability-induced retrieval hypothesis: contextual variability was beneficial when training problems were spaced in time, presumably inducing effortful retrieval (akin to a closed-book test), but not when training problems and their solutions remained visible to help solve subsequent problems (akin to an open-book test). Together these findings suggest that contextual variability can enhance transfer not only via structural alignment of to-be-studied examples, but also via increased retrieval demands across successive to-be-solved problems.

It is important to emphasize that this hypothesis is not entirely new; others have noted that contextual variability is one of several ways to impose retrieval demands during training and thereby enhance subsequent retention (e.g., Butler et al., 2017; Cuddy & Jacoby, 1982; Glass, 2009; Jacoby, 1978). Nonetheless, the findings of Chapter 2 are salient in that it supports this hypothesis using a more complex transfer task rather than a simpler task involving word stimuli.

How, then, might the variability-induced retrieval hypothesis inform educational design? A strong interpretation of the results in Chapter 2 suggests that, in contrast to presenting dissimilar problems side-by-side and having learners compare them, there may be situations where it is wiser to present one of the problems during initial learning and the other after a significant delay. Yet I am reluctant to fully embrace this conclusion because: (i) the experiments in Chapter 2 involved learners solving training problems in an unguided fashion, unlike more educationally relevant scenarios where the instructor usually prompts some type of explicit comparison; (ii) there is a sizable literature on the benefits of being explicitly prompted to

compare problems posed in different contexts (see Alfieri et al., 2013); and (iii) the benefits of comparing problems in different contexts likely depends on how context is operationally defined[11], the nature of the problem, and the type of principle being learned. Perhaps the more conservative implication is that practice testing techniques might be augmented by varying the contexts in which test problems are presented. For example, when designing multiple test questions to enhance understanding of a given principle, instructors could ensure that the questions differ in their surface features (e.g., the objects described in the problem, the question being asked) such that retrieval of relevant prior examples becomes nontrivial.

How Exactly Does 'Effortful Retrieval' Promote Learning?

But what exactly is meant by more 'effortful' or 'difficult' retrieval attempts being beneficial for learning? As discussed in Chapter 1, basic memory experiments reliably show that the effort or difficulty of a retrieval attempt correlates positively with subsequent retention (e.g., Carpenter, 2009; 2011; Pyc & Rawson, 2009; for meta-analysis, see Rowland, 2014). But these terms are merely labels that we use to describe a learner's performance based on either intuition or other proxies like longer response times, lower accuracy, or greater perceived effort. In other words, explanations related to effort or difficulty are descriptive rather than process-oriented, and so do not lend themselves to testable predictions about underlying mechanisms. A more fruitful approach might be to consider hypotheses that explain why effortful retrieval attempts benefit learning at a mechanistic level. Two prominent mechanistic explanations for the benefits of effortful retrieval are discussed below.

Elaborative Retrieval

The elaborative retrieval hypothesis posits that effortful retrieval attempts initiate a search process that activates similar information in memory, enriching the memory trace so it can be accessed via additional retrieval cues in the future (e.g., Carpenter, 2009; 2011; Carpenter & Yeung, 2017; Rawson et al., 2015). To illustrate this point, Carpenter (2011) had participants study cue–target pairs (e.g., Donor–Heart), then either engage in retrieval practice via cued recall

[11] Note how context can be framed in practically infinite different ways: the specific objects described in a problem, the type of question being posed by a problem, the learner's physical environment, the mental state of the learner, the time of day in which the problem is administered, and so on. For this reason I think the true challenge lies in understanding *what* forms of contextual variability are beneficial depending on the specific learning task at hand.

(e.g., Donor–_____) or study the intact pairs again (e.g., Donor–Heart). Critically, the pairs were constructed so that a specific semantic mediator (e.g., Blood) had a high association strength with the cue word. A final recognition test revealed higher false alarm rates for these semantic mediator words from participants who had undergone cued recall relative to the study-only group, consistent with the idea that retrieval of the target word (Heart) activated related information (Blood) in memory. These findings suggest that effortful retrieval attempts are beneficial because they trigger activation of related information, rendering the studied information more accessible through the newly incorporated retrieval cues.

However, the elaborative retrieval hypothesis has recently fallen under scrutiny because of contradictory findings, most notably when retrieval practice is directly compared with elaborative encoding manipulations. For example, Lehman et al. (2014) had participants study four lists of words in sequence. After each list, the retrieval practice group engaged in free recall for the just-shown list, whereas the elaboration group received the words from the just-shown list and typed the first two words that came to mind. Critically, a final free recall test for all the lists showed far superior performance for the retrieval practice group than the elaboration group. Such findings and others like them (Lehman & Karpicke, 2016; also see arguments about cue overload by Watkins & Watkins, 1975; Wixted & Rohrer, 1993) suggest that retrieval practice exerts its benefits through a mechanism other than semantic elaboration.

The Multiple Trace View and Release from Interference

A lesser-known hypothesis focuses on retrieval practice as a means to segment learning events into more discrete memory representations, thereby overcoming the type of interference that arises after studying similar sets of information close in time. Resembling multiple trace theories (e.g., Moscovitch et al., 2005; Nadel et al., 2000), this hypothesis suggests that effortful retrieval attempts induce contextual changes that better segment or ‘chunk’ the studied information in memory. Said segmenting of the information is then thought to improve subsequent access because the learner’s subsequent retrieval attempts can be constrained to a smaller portion of what they learned, mitigating interference from the other studied information that competes for access at the time of retrieval (e.g., Szpunar et al., 2008; for review, see Yang et al., 2018). By this view, effortful retrieval attempts promote segmentation of the studied information as if more distinct memory traces are formed, imposing a richer higher-order structure that can later be used to home in on the desired information.

This multiple trace view also seems consistent with the key finding in Chapter 2. In the case of the adjacent format, because training problems for a given principle were solved with minimal retrieval demands, one might expect this type of learning to produce one larger, unsegmented memory trace. Likewise, in the case of the spaced format when training problems for a given principle had highly similar surface features, the close match in context across training problems may have stifled the need for effortful retrieval, again producing one unsegmented memory trace. This logic would explain why far transfer performance was poor following the adjacent format regardless of contextual variability, and also in the spaced format in the absence of contextual variability. In contrast, solving training problems in the spaced condition with a high degree of contextual variability ought to impose the greatest retrieval demands, and thus the highest probability of forming distinct traces for each training problem. This would explain why far transfer performance was only improved when contextual variability was paired with the spaced format. This multiple trace interpretation remains speculative, however, as the experiments in Chapter 2 were not designed to test this hypothesis.

Although the multiple trace view for the benefits of retrieval practice has recently garnered much attention from researchers, most of the evidence comes from basic memory experiments involving the recall of word stimuli (e.g., Bäuml & Kliegl, 2013; Szpunar et al., 2008), text passages (e.g., Wissman et al., 2011), face–name associations (e.g.,. Weinstein et al., 2011), or the categorization of obscure images (e.g., Lee & Ahn, 2018; Yang & Shanks, 2018). It therefore remains unclear whether the same logic extends to more complex learning tasks reminiscent of those encountered in higher education (see Yang et al., 2018). Nonetheless, because the negative effects of interference are well-documented in a range of different learning tasks[12] (e.g., Dillon et al., 1973; Marton et al., 2014; Weinstein et al., 2011), I contend that this mechanistic explanation deserves more attention to explain when and why retrieval demands are beneficial for more complex forms of learning involving problem solving and transfer. I am currently conducting other experiments with this goal in mind.

[12] Speaking further to its robustness, interference also reliably impairs learning in many non-human animals, including rodents (e.g., Dunnet et al., 1990), monkeys (e.g., Jitsumori et al., 1988), and birds (e.g., Grant, 1975).

A Final Note on the Importance of Retrieval Demands

It stands to reason that retrieval practice—or as I have framed it here, introducing greater retrieval demands during training—may be a key ingredient for the type of durable and flexible learning revered by educators. As discussed in the section above, however, mechanistic explanations for the benefits of retrieval practice are restrained to simpler learning tasks that do not closely resemble those encountered in authentic educational settings. Only by further elucidating these mechanisms in relation to more complex learning tasks will we be able to offer concrete advice about structuring learning to optimize transfer.

The Correlational Approach

Little Evidence for Validity of SPQ Scores

The primary goal of Chapters 3 and 4 was to assess the predictive validity of the SPQ using transfer-oriented performance measures. A secondary goal of Chapter 3 was to assess whether SPQ scores could predict transfer in a context-independent manner as if measuring a general trait (i.e., without responses being tied to a specific learning context). The data in Chapter 3 showed that SPQ deep approach scores did not predict transfer in a context-general manner during a laboratory task, nor in a context-specific manner during a university engineering course. For the latter classroom study, however, transfer was operationally defined based solely on the perceptions of instructors. I therefore adopted a novel student-centered approach in Chapter 4 whereby transfer measures were operationally defined based on the perceptions of students. Again, SPQ scores did not, or else very weakly, predicted performance on exam questions that students collectively perceived to demand transfer. Given the large sample sizes in these studies and the replication of similar results across three separate university courses, these data arguably call into question the utility of the SPQ in educational practice.

It is critical to note that my colleagues and I are not alone in challenging the validity of the SPQ or other similar inventories. From a theoretical standpoint, the central criticism is that conceptions of deep learning are ill-defined, with minimal focus on specific types of information processing and the corresponding outcome measures that serve as evidence of said processing having occurred (e.g., Dinsmore & Alexander, 2012; Howie & Bagnall, 2013). From an empirical standpoint, correlations between SPQ deep approach scores and course performance are usually below r = .20 (e.g., Cantwell & Moore, 1998; McManus et al. 1998; Snelgrove &

Slater, 2003; Zeegers, 2001; for meta-analysis, see Richardson et al., 2012) or else statistically insignificant (e.g., Gijbels et al., 2005; Groves, 2005; Wilding & Andrews, 2006). Indeed, it was these theoretical and empirical concerns that led us to emphasize transfer-oriented outcomes in the first place. Yet these concerns seem to have gone largely ignored: the SPQ and other similar inventories are still routinely adopted under the assumption that they capture meaningful constructs associated with student learning.

The rest of this chapter is therefore dedicated to two questions: why are deep learning inventories like the SPQ still trumpeted as tools to improve education if there is so little evidence for their predictive validity? And is there any hope for such inventories moving forward?

A Tenuous Link to Levels of Processing

Part of the SPQ's appeal stems from its similarity to the influential levels of processing framework (e.g., Craik & Lockhart, 1972). Briefly, this framework emphasizes 'depth' of encoding such that stimuli are better retained when processed with respect to semantics rather than superficial physical features. For example, participants might process words at a 'deep' level by being oriented toward meaning (e.g., by classifying the referent of each word as animate or not) or at a 'shallow' level by being oriented toward orthography (e.g., by counting the number of vowels in each word). Semantic processing reliably improves performance on subsequent memory tests more so than shallower forms of processing (see Craik, 2002; Roediger, 2008). Although mainly developed to explain memory phenomena with simple word stimuli, the levels of processing framework is also routinely cited as theoretical justification for deep learning inventories (e.g., Entwistle & McCune, 2004; Ross et al., 2006; Dinsmore & Alexander, 2012). The SPQ is thus viewed as a tool to 'scale up' the levels of processing logic by characterizing how students approach more complex and realistic learning tasks.

But the link between levels of processing and the SPQ and is overstated. First, the levels of processing framework itself has been criticized for vague indices of 'depth' and circularity in logic (e.g., Craik, 2002; Eysenck, 1978). That is, enhanced retention following some intervention can be attributed to semantic processing based on an intuitive notion of task demands, but this does not explain *why* said processing enhanced retention[13]. Second, levels of processing studies

[13] It also does not explain why different forms of semantic encoding can elicit different memory outcomes (for a more detailed critique along these lines, see Jacoby et al., 1979)

usually rely on artificial stimuli (e.g., words) and memory tests (e.g., recognition or verbatim recall), so their relevance to more complex learning is not self-evident. After all, what does it mean to process a diagram of the human heart or a textbook chapter at a 'semantic' level? And besides tests of recall, what outcome measures constitute evidence of said semantic processing having occurred? These questions are not easy to answer. Third, most SPQ items appear to reflect general motivation rather than specific types of information processing. In fact, the items are written in a way so that it becomes obvious some responses are 'good' and others 'bad' (also see Howie & Bagnall, 2013, p. 394). The SPQ might therefore capture a student's perceptions of their general motivation or aptitude instead of the forms of processing they tend to adopt. In short, couching the SPQ in terms of levels of processing seems simplistic at best.

Reliance on Introspection

There is also something to be said about these inventories relying exclusively on self-reports. I will not belabour the point here because others have discussed the limitations of introspection at greater length (e.g., Dunlosky & Rawson, 2012; Dunning et al., 2003; Eva & Regehr, 2005; Nisbett & Wilson, 1977). It is worth noting, however, that the issue of inaccurate introspection is likely to be exacerbated when implementing these inventories because they: (i) probe for perceptions of abstract things like motivation rather than performance on a concrete task; (ii) are typically administered only once during an entire course, meaning that students must introspect about their attitudes and behaviours over a long timespan; and (iii) pertain to attitudes and study habits that a student may have adopted relatively routinely and automatically during their studies, perhaps impairing introspection (also see Kellog, 1982). Such shortcomings underscore the need to be skeptical of how these self-report measures are interpreted.

A Fool's Errand?

Beyond the SPQ, the ultimate question pertains to the feasibility of any such tool designed to predict transfer outcomes via self-reported individual differences. The challenge is one of context specificity, because it seems misguided to coarsely measure a student's perceptions of their learning across an entire course and expect this to predict performance on a wide array of transfer tasks, each requiring a unique set of knowledge or skills. As argued by Salomon and Perkins (1989), transfer should not be viewed as a unitary phenomenon, but rather in terms of "what" knowledge or skill is actually being transferred. So, by neglecting context

specificity, the pursuit of a tool that predicts a diverse range of transfer outcomes by capturing student perceptions over a large time scale may prove to be a fool's errand.

Extending this argument about context specificity, future work might find greater evidence for the validity of these self-report inventories through repeated administrations. For example, instead of administering the SPQ only once during a course and examining overall exam performance, it may be worthwhile to administer it after every unit and then selectively examine performance on transfer-oriented assessments that correspond to each of those units. In this way, approach scores obtained via the SPQ are more closely tied to specific sets of content and how they were learned, mitigating concerns about context specificity and the students having to introspect on behaviour over long timeframes. This idea awaits further study[14].

A Final Note on The Construct of Deep Learning

In sum, because the SPQ and the larger construct of deep learning continue to dominate the education literature, the story emerging here is sobering. As argued by Roediger (2013):

> The field of education seems particularly susceptible to the allure of plausible but untested ideas and fads (especially ones that are lucrative for their inventors) . . . . One could write an interesting history of ideas based on either plausible theory or somewhat flimsy research . . . and once an idea takes hold, it is hard to root out." (p. 2)

Along the same lines, I argue that inventories meant to quantify deep learning—and the construct of deep learning itself—fall under this description. Deep learning certainly seems like an intuitive and useful idea to educators, even if only to express a desire for learning outcomes associated with application of knowledge, problem-solving, and transfer (Halpern & Hakel, 2003). Informal discussions about deep learning are also beneficial in that they prompt educators to reflect on their teaching practices. Nevertheless, as scientists we must step back from all the rhetoric and ask whether the various and dizzying descriptions of deep learning add to our understanding of transfer in a meaningful way. Theory is only valuable insofar as it produces clear and testable predictions that align with empirical data—yet the construct of deep learning

[14] Note how claims of context specificity further obscure the SPQ's utility: If its approach scores fail to predict transfer performance, it is always possible to argue that the transfer task did not align with the particular learning context that gave rise to the SPQ scores. At one extreme, this would mean the validity of the SPQ hinges upon it being completed for every individual learning task encountered by students!

does not convincingly meet these criteria. Maybe, then, it is time to abandon these learning inventories until we reach a deeper understanding of what we mean by deep learning and how to properly measure whether it occurred or not.

Concluding Remarks

I laid out the overall structure of this thesis in Chapter 1, drawing upon Cronbach's (1957) distinction between two approaches in psychology. With respect to studies on transfer, the work discussed here suggests that more success has been found through the experimental approach than the correlational approach. But, like Cronbach, I am now left wondering about the marriage of these two approaches. My hope is that future correlational research on transfer will begin to encompass the findings that have arisen through experimentation. That is, instead of focusing on broad constructs like motivation, perhaps self-report inventories can be geared toward the specific mental operations a student tends to adopt. For example, one could devise a questionnaire that asks students to what extent they compare dissimilar examples of same principle, engage in retrieval practice, distribute their learning over time, or introduce contextual variability in their studying. This hybrid approach seems timely. After all, while controlled experiments have yielded many promising findings thus far, at some point we must acknowledge that the complexity of human learning requires us to embrace both approaches in tandem.

List of References

Agarwal, P. K., Bain, P. M., & Chamberlain, R. W. (2012). The value of applied research: Retrieval practice improves classroom learning and recommendations from a teacher, a principal, and a scientist. *Educational Psychology Review*, *24*(3), 437–448.

Alfieri, L., Nokes-Malach, T. J., & Schunn, C. D. (2013). Learning through case comparisons: A meta-analytic review. *Educational Psychologist*, *48*(2), 87–113.

Baeten, M., Kyndt, E., Struyven, K., & Dochy, F. (2010). Using student-centred learning environments to stimulate deep approaches to learning: Factors encouraging or discouraging their effectiveness. *Educational Research Review, 5*(3), 243–260.

Bartlett, F. C. (1932). Remembering: A study in experimental and social psychology. Cambridge: Cambridge University Press

Bassok, M., & Holyoak, K. J. (1989). Interdomain transfer between isomorphic topics in algebra and physics. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(1), 153–166.

Bäuml, K. H. T., & Kliegl, O. (2013). The critical role of retrieval processes in release from proactive interference. *Journal of Memory and Language*, *68*(1), 39–53.

Biggs, J. (1993). What do inventories of students' learning processes really measure? A theoretical review and clarification. *British Journal of Educational Psychology, 63*(1), 3–19.

Biggs, J. B. (1987a). *Student Approaches to Learning and Studying.* Hawthorn, Australia: Australian Council for Educational Research Ltd.

Biggs, J., Kember, D., & Leung, D. Y. (2001). The revised two-factor study process questionnaire: R-SPQ-2F. *British Journal of Educational Psychology*, *71*(1), 133–149.

Billing, D. (2007). Teaching for transfer of core/key skills in higher education: Cognitive skills. *Higher Education*, *53*(4), 483–516.

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. *Psychology and the real world: Essays illustrating fundamental contributions to society*, *2*(59–68).

Bouckenooghe, D., Cools, E., De Clercq, D., Vanderheyden, K., & Fatima, T. (2016). Exploring the impact of cognitive style profiles on different learning approaches: Empirical evidence for adopting a person-centered perspective. *Learning and Individual Differences*, *51*, 299–306.

Bransford, J. D., & Schwartz, D. L. (1999). Chapter 3: Rethinking transfer: A simple proposal with multiple implications. *Review of Research in Education*, *24*(1), 61–100.

Brooks, L. R., & Hannah, S. D. (2006). Instantiated features and the use of" rules.". *Journal of Experimental Psychology: General*, *135*(2), 133–151.

Brooks, L. R., & Vokey, J. R. (1991). Abstract analogies and abstracted grammars: Comments on Reber (1989) and Mathews et al. (1989). *Journal of Experimental Psychology: General*, 120, 316–323.

Brown, A. L., Kane, M. J., & Echols, C. H. (1986). Young children's mental models determine analogical transfer across problems with a common goal structure. *Cognitive Development*, *1*(2), 103–121.

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(5), 1118–1133.

Butler, A. C., & Roediger III, H. L. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology*, *19*(4-5), 514–527.

Butler, A. C., Black-Maier, A. C., Raley, N. D., & Marsh, E. J. (2017). Retrieving and applying knowledge to different examples promotes transfer of learning. *Journal of Experimental Psychology: Applied*, *23*(4), 433–446.

Cantwell H. & Moore P.J. (1998) Relationships among control beliefs, approaches to learning and the academic achievement of final year nurses. *The Alberta Journal of Educational Research*, *1*, 98–102.

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: the benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(6), 1563–1569.

Carpenter, S. K. (2011). Semantic information activated during retrieval contributes to later retention: Support for the mediator effectiveness hypothesis of the testing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(6), 1547–1552.

Carpenter, S. K. (2012). Testing enhances the transfer of learning. *Current Directions in Psychological Science*, *21*(5), 279–283.

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition*, *34*(2), 268–276.

Carpenter, S. K., & Yeung, K. L. (2017). The role of mediator strength in learning from retrieval. *Journal of Memory and Language*, *92*, 128–141.

Catrambone, R. (1996). Generalizing solution procedures learned from examples. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(4), 1020–1031.

Catrambone, R., & Holyoak, K. J. (1989). Overcoming contextual limitations on problem-solving transfer. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(6), 1147–1156.

Chamorro-Premuzic, T., & Furnham, A. (2009). Mainly Openness: The relationship between the Big Five personality traits and learning approaches. *Learning and Individual Differences*, *19*(4), 524–529.

Chamorro-Premuzic, T., Furnham, A., & Lewis, M. (2007). Personality and approaches to learning predict preference for different teaching methods. *Learning and Individual Differences*, *17*(3), 241–250.

Chan, J. C., McDermott, K. B., & Roediger III, H. L. (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General*, *135*(4), 553–571.

Chernobilsky, E., DaCosta, M. C., & Hmelo-Silver, C. E. (2004). Learning to talk the educational psychology talk through a problem-based course. *Instructional Science*, *32*(4), 319–356.

Choy, J. L. F., O'Grady, G., & Rotgans, J. I. (2012). Is the Study Process Questionnaire (SPQ) a good predictor of academic achievement? Examining the mediating role of achievement-related classroom behaviours. *Instructional Science*, *40*(1), 159–172.

Christie, S., & Gentner, D. (2010). Where hypotheses come from: Learning new relations by structural alignment. *Journal of Cognition and Development*, *11*(3), 356–373.

Craik, F. I. (2002). Levels of processing: Past, present... and future? *Memory*, *10*(5-6), 305–318.

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 671–684.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, *12*(11), 671–684.

Cuddy, L. J., & Jacoby, L. L. (1982). When forgetting helps memory: An analysis of repetition effects. *Journal of Verbal Learning and Verbal Behavior*, *21*(4), 451–467.

Day, S. B., & Gentner, D. (2007). Nonintentional analogical inference in text comprehension. *Memory & Cognition*, *35*(1), 39–49.

Day, S. B., & Goldstone, R. L. (2011). Analogical transfer from a simulated physical system. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(3), 551–567.

Day, S. B., & Goldstone, R. L. (2012). The import of knowledge export: Connecting findings and theories of transfer of learning. *Educational Psychologist*, *47*(3), 153–176.

Delaney, P. F., Verkoeijen, P. P., & Spirgel, A. (2010). Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. In *Psychology of Learning and Motivation* (Vol. 53, pp. 63–147). Academic Press.

Dillon, R. F., McCormack, P. D., Petrusic, W. M., Cook, G. M., & Lafleur, L. (1973). Release from proactive interference in compound and coordinate bilinguals. *Bulletin of the Psychonomic Society*, *2*(5), 293–294.

Dinsmore, D. L., & Alexander, P. A. (2012). A critical discussion of deep and surface processing: What it means, how it is measured, the role of context, and model specification. *Educational Psychology Review*, *24*(4), 499–567.

Duff, A., Boyle, E., Dunleavy, K., & Ferguson, J. (2004). The relationship between personality, approach to learning and academic performance. *Personality and Individual Differences*, *36*(8), 1907–1920.

Duncker, K. (1945). On problem-solving. *Psychological Monographs*, *270*, i.

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learning and Instruction*, *22*(4), 271–280.

Dunnett, S. B., Martel, F. L., & Iversen, S. D. (1990). Proactive interference effects on short-term memory in rats: II. Effects in young and aged rats. *Behavioral Neuroscience*, *104*(5), 666–670.

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science*, *12*(3), 83–87.

Eglington, L. G., & Kang, S. H. (2018). Retrieval practice benefits deductive inference. *Educational Psychology Review*, *30*(1), 215–228.

Entwistle, N., & McCune, V. (2004). The conceptual bases of study strategy inventories. *Educational Psychology Review*, *16*(4), 325–345.

Entwistle, N., Hanley, M., & Hounsell, D. (1979). Identifying distinctive approaches to studying. *Higher Education*, *8*(4), 365–380.

Eva, K. W., & Regehr, G. (2005). Self-assessment in the health professions: A reformulation and research agenda. *Academic Medicine*, *80*(10), 46–54.

Eva, K. W., Neville, A. J., & Norman, G. R. (1998). Exploring the etiology of content specificity: Factors influencing analogic transfer and problem solving. *Academic Medicine*, *73*(10), 1–5.

Eysenck, M. W. (1978). Levels of processing: A critique. *British Journal of Psychology*, *69*(2), 157–169.

Frey, R. F., Cahill, M. J., & McDaniel, M. A. (2017). Students' concept-building approaches: A novel predictor of success in chemistry courses. *Journal of Chemical Education*, *94*(9), 1185–1194.

Fyfe, E. R., McNeil, N. M., Son, J. Y., & Goldstone, R. L. (2014). Concreteness fading in mathematics and science instruction: A systematic review. *Educational Psychology Review*, *26*(1), 9–25.

Gartman, L. M., & Johnson, N. F. (1972). Massed versus distributed repetition of homographs: A test of the differential-encoding hypothesis. *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 801–808.

Gentner, D., & Smith, L. (2012). Analogical reasoning. *Encyclopedia of Human Behavior* (2nd Ed.), *1*, 130–136.

Gentner, D., & Toupin, C. (1986). Systematicity and surface similarity in the development of analogy. *Cognitive Science*, *10*(3), 277–300.

Gentner, D., Loewenstein, J., & Thompson, L. (2003). Learning and transfer: A general role for analogical encoding. *Journal of educational psychology*, *95*(2), 393–408.

Gick, M. L., & Holyoak. K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, *15*, 1–38.

Gijbels, D., Van de Watering, G., Dochy, F., & Van den Bossche, P. (2005). The relationship between students' approaches to learning and the assessment of learning outcomes. *European Journal of Psychology of Education*, *20*(4), 327–341.

Glass, A. L. (2009). The effect of distributed questioning with varied examples on exam performance on inference questions. *Educational Psychology*, *29*, 831–848.

Glenberg, A. M. (1979). Component-levels theory of the effects of spacing of repetitions on recall and recognition. *Memory & Cognition*, *7*(2), 95–112.

Gobet, F., & Simon, H. A. (1996). Recall of random and distorted chess positions: Implications for the theory of expertise. *Memory & Cognition*, *24*(4), 493–503.

Goldstone, R. L., & Sakamoto, Y. (2003). The transfer of abstract principles governing complex adaptive systems. *Cognitive Psychology*, *46*(4), 414–466.

Grant, D. S. (1975). Proactive interference in pigeon short-term memory. *Journal of Experimental Psychology: Animal Behavior Processes*, *1*(3), 207–220.

Greene, R. L., & Stillwell, A. M. (1995). Effects of encoding variability and spacing on frequency discrimination. *Journal of Memory and Language*, *34*(4), 468–476.

Groves, M. (2005). Problem-based learning and learning approach: Is there a relationship? *Advances in Health Sciences Education*, *10*(4), 315–326.

Halpern, D. F., & Hakel, M. D. (2003). Applying the science of learning to the university and beyond: Teaching for long-term retention and transfer. *Change: The Magazine of Higher Learning*, *35*(4), 36–41.

Halpern, D. F., Hansen, C., & Riefer, D. (1990). Analogies as an aid to understanding and memory. *Journal of Educational Psychology*, *82*(2), 298–305.

Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers*, *16*(2), 96–101.

Holyoak, K. J., & Koh, K. (1987). Surface and structural similarity in analogical transfer. *Memory & Cognition*, *15*(4), 332–340.

Holyoak, K. J., & Koh, K. (1987). Surface and structural similarity in analogical transfer. *Memory & Cognition*, *15*(4), 332–340.

Howie, P., & Bagnall, R. (2013). A critique of the deep and surface approaches to learning model. *Teaching in Higher Education*, *18*(4), 389–400.

Jacoby, L. L. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal of Verbal Learning and Verbal Behavior*, *17*(6), 649–667.

Jacoby, L. L., Craik, F. I, & Begg, I (1979). Effects of decision difficulty on recognition and recall. *Journal of Verbal Learning and Verbal Behavior*, *18*(5), 585–600.

Jacoby, L. L., Wahlheim, C. N., & Coane, J. H. (2010). Test-enhanced learning of natural concepts: Effects on recognition memory, classification, and metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(6), 1441–1451.

James, W. (1899) *On some of life's ideals*. New York: Henry

Jamieson, R. K., Crump, M. J., & Hannah, S. D. (2012). An instance theory of associative learning. *Learning & Behavior*, *40*(1), 61–82.

Jarus, T., & Goverover, Y. (1999). Effects of contextual interference and age on acquisition, retention, and transfer of motor skill. *Perceptual and Motor Skills*, *88*(2), 437–447.

Jitsumori, M., Wright, A. A., & Cook, R. G. (1988). Long-term proactive interference and novelty enhancement effects in monkey list memory. *Journal of Experimental Psychology: Animal Behavior Processes, 14*(2), 146–154.

Kaminski, J. A., Sloutsky, V. M., & Heckler, A. F. (2013). The cost of concreteness: The effect of nonessential information on analogical transfer. *Journal of Experimental Psychology: Applied, 19*(1), 14–29.

Karpicke, J. D., & Aue, W. R. (2015). The testing effect is alive and well with complex materials. *Educational Psychology Review, 27*(2), 317–326.

Karpicke, J. D., & Zaromb, F. M. (2010). Retrieval mode distinguishes the testing effect from the generation effect. *Journal of Memory and Language, 62*(3), 227–239.

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning: An episodic context account. In *Psychology of Learning and Motivation* (Vol. 61, pp. 237–284). Academic Press.

Kellogg, R. T. (1982). When can we introspect accurately about mental processes? *Memory & Cognition, 10*(2), 141–144.

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction"? *Psychological Science, 19*(6), 585–592.

Kornell, N., & Vaughn, K. E. (2016). How retrieval attempts affect learning: A review and synthesis. In *Psychology of Learning and Motivation* (Vol. 65, pp. 183–215). Academic Press.

Kostic, B., Cleary, A. M., Severin, K., & Miller, S. W. (2010). Detecting analogical resemblance without retrieving the source analogy. *Psychonomic Bulletin & Review, 17*(3), 405–411.

Kotovsky, L., & Gentner, D. (1996). Comparison and categorization in the development of relational similarity. *Child Development, 67*(6), 2797–2822.

Laird, T. F., Seifert, T. A., Pascarella, E. T., Mayhew, M. J., & Blaich, C. F. (2014). Deeply affecting first-year students' thinking: Deep approaches to learning and three dimensions of cognitive development. *The Journal of Higher Education, 85*(3), 402–432.

Landin, D. K., Hebert, E. P., & Fairweather, M. (1993). The effects of variable practice on the performance of a basketball skill. *Research Quarterly for Exercise and Sport*, *64*(2), 232–237.

Larsen, D. P., Butler, A. C., Lawson, A. L., & Roediger, H. L. (2013). The importance of seeing the patient: test-enhanced learning with standardized patients and written tests improves clinical application of knowledge. *Advances in Health Sciences Education, 18*(3), 409–425.

Lee, H. S., & Ahn, D. (2018). Testing prepares students to learn better: The forward effect of testing in category learning. *Journal of Educational Psychology*, *110*(2), 203–217.

Lehman, M., & Karpicke, J. D. (2016). Elaborative retrieval: Do semantic mediators improve memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(10), 1573–1591.

Lehman, M., Smith, M. A., & Karpicke, J. D. (2014). Toward an episodic context account of retrieval-based learning: Dissociating retrieval practice and elaboration. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(6), 1787–1794.

Lewis, F. C. (1905). A study in formal discipline. *The School Review*, *13*(4), 281–292.

Loewenstein, J., & Gentner, D. (2005). Relational language and the development of relational mapping. *Cognitive Psychology*, *50*(4), 315–353.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, *95*(4), 492–527.

Logan, G. D., Taylor, S. E., & Etherton, J. L. (1996). Attention in the acquisition and expression of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(3), 620–638.

Markman, A. B., & Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive Psychology*, *25*(4), 431–467.

Marton, F., & Säljö, R. (1976). On qualitative differences in learning: I—Outcome and process. *British Journal of Educational Psychology*, *46*(1), 4–11.

Marton, K., Campanelli, L., Eichorn, N., Scheuer, J., & Yoon, J. (2014). Information processing and proactive interference in children with and without specific language impairment. *Journal of Speech, Language, and Hearing Research*, 57(1), 106–119.

Mayer, R. E. (2008). Applying the science of learning: Evidence-based principles for the design of multimedia instruction. *American Psychologist*, *63*(8), 760–769.

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, *19*(4-5), 494–513.

McDaniel, M. A., Cahill, M. J., Frey, R. F., Rauch, M., Doele, J., Ruvolo, D., & Daschbach, M. M. (2018). Individual differences in learning exemplars versus abstracting rules: Associations with exam performance in college science. *Journal of Applied Research in Memory and Cognition*, *7*(2), 241–251.

McDaniel, M. A., Cahill, M. J., Robbins, M., & Wiener, C. (2014). Individual differences in learning and transfer: Stable tendencies for learning exemplars versus abstracting rules. *Journal of Experimental Psychology: General*, *143*(2), 668–693.

McManus, I. C., Richards, P., Winder, B. C., & Sproston, K. A. (1998). Clinical experience, performance in final examinations, and learning style in medical students: Prospective study. *British Medical Journal*, *316*(7128), 345–350.

Medin, D. L., & Ross, B. H. (1989). *The specific character of abstract thought: Categorization, problem solving, and induction.* In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence, Vol. 5* (p. 189–223). Lawrence Erlbaum Associates, Inc.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207–238.

Melton, A. W. (1970). The situation with respect to the spacing of repetitions and memory. *Journal of Verbal Learning and Verbal Behavior*, *9*(5), 596–606.

Miyatsu, T., Gouravajhala, R., Nosofsky, R. M., & McDaniel, M. A. (2019). Feature highlighting enhances learning of a complex natural-science category. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(1), 1–16.

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*(5), 519–533.

Moscovitch, M., Rosenbaum, R. S., Gilboa, A., Addis, D. R., Westmacott, R., Grady, C., McAndrews, M. P., Levine., B., Black, S., Winocur, G., & Nadel, L. (2005). Functional neuroanatomy of remote episodic, semantic and spatial memory: A unified account based on multiple trace theory. *Journal of Anatomy*, *207*(1), 35–66.

Nadel, L., Samsonovich, A., Ryan, L., & Moscovitch, M. (2000). Multiple trace theory of human memory: Computational, neuroimaging, and neuropsychological results. *Hippocampus*, *10*(4), 352–368.

Needham, D. R., & Begg, I. M. (1991). Problem-oriented training promotes spontaneous analogical transfer: Memory-oriented training promotes memory for training. *Memory & Cognition*, *19*(6), 543–557.

Newell, K. M., & Shapiro, D. C. (1976). Variability of practice and transfer of training: Some evidence toward a schema view of motor learning. *Journal of Motor Behavior*, *8*(3), 233–243.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231–259.

Norman, G. (2009). Teaching basic science to optimize transfer. *Medical Teacher*, *31*(9), 807–811.

Pask, G., & Scott, B. C. E. (1972). Learning strategies and individual competence. *International Journal of Man-Machine Studies*, *4*(3), 217–253.

Postman, L., & Knecht, K. (1983). Encoding variability and retention. *Journal of Verbal Learning and Verbal Behavior*, *22*(2), 133–152.

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, *60*(4), 437–447.

Quilici, J. L., & Mayer, R. E. (2002). Teaching students to recognize structural similarities between statistics word problems. *Applied Cognitive Psychology*, *16*(3), 325–342.

Rawson, K. A. (2015). The status of the testing effect for complex materials: Still a winner. *Educational Psychology Review*, *27*(2), 327–331.

Rawson, K. A., Vaughn, K. E., & Carpenter, S. K. (2015). Does the benefit of testing depend on lag, and if so, why? Evaluating the elaborative retrieval hypothesis. *Memory & Cognition*, *43*(4), 619–633.

Reeves, L., & Weisberg, R. W. (1994). The role of content and abstract information in analogical transfer. *Psychological Bulletin*, *115*(3), 381–400.

Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin*, *138*(2), 353–387.

Richland, L. E., & McDonough, I. M. (2010). Learning by analogy: Discriminating between potential analogs. *Contemporary Educational Psychology*, *35*(1), 28–43.

Rittle-Johnson, B., & Star, J. R. (2009). Compared with what? The effects of different comparisons on conceptual knowledge and procedural flexibility for equation solving. *Journal of Educational Psychology*, *101*(3), 529–544.

Roediger III, H. L. (2013). Applying cognitive psychology to education: Translational educational science. *Psychological Science in the Public Interest*, *14*(1), 1–3.

Roediger III, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20–27.

Roediger, III, H. L. (2008). Relativity of remembering: Why the laws of memory vanished. *Annual Review of Psychology*, *59*, 225–254.

Rohrer, D., Taylor, K., & Sholar, B. (2010). Tests enhance the transfer of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(1), 233–239.

Ross, B. H. (1987). This is like that: The use of earlier problems and the separation of similarity effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(4), 629–639.

Ross, B. H. (1989). Distinguishing types of superficial similarities: Different effects on the access and use of earlier problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(3), 456–468.

Ross, B. H., & Kennedy, P. T. (1990). Generalizing from the use of earlier examples in problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(1), 42–55.

Ross, M. E., Green, S. B., Salisbury-Glennon, J. D., & Tollefson, N. (2006). College students' study strategies as a function of testing: An investigation into metacognitive self-regulation. *Innovative Higher Education*, *30*(5), 361–375.

Rowland, C. A. (2014). The effect of testing versus restudy on retention: a meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432–1463.

Salomon, G., & Perkins, D. N. (1989). Rocky roads to transfer: Rethinking mechanism of a neglected phenomenon. *Educational Psychologist*, *24*(2), 113–142.

Sandberg, J., & Barnard, Y. (1997). Deep learning is difficult. *Instructional Science*, *25*(1), 15–36.

Schunn, C. D., & Dunbar, K. (1996). Priming, analogy, and awareness in complex reasoning. *Memory & Cognition*, *24*(3), 271–284.

Scouller, K. M., & Prosser, M. (1994). Students' experiences in studying for multiple choice question examinations. *Studies in Higher Education*, *19*(3), 267–279.

Slamecka, N. J., & Barlow, W. (1979). The role of semantic and surface features in word repetition effects. *Journal of Verbal Learning and Verbal Behavior*, *18*(5), 617–627.

Smith, S. M., & Vela, E. (2001). Environmental context-dependent memory: A review and meta-analysis. *Psychonomic Bulletin & Review*, *8*(2), 203–220.

Struyven, K., Dochy, F., Janssens, S., & Gielen, S. (2006). On the dynamics of students' approaches to learning: The effects of the teaching/learning environment. *Learning and Instruction*, *16*(4), 279–294.

Szpunar, K. K., & McDermott, K. B. (2008). Episodic future thought and its relation to remembering: Evidence from ratings of subjective experience. *Consciousness and Cognition*, *17*(1), 330–334.

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, *80*(5), 352–373

Van Gog, T., & Sweller, J. (2015). Not new, but nearly forgotten: The testing effect decreases or even disappears as the complexity of learning materials increases. *Educational Psychology Review*, *27*(2), 247–264.

Verkoeijen, P. P., Rikers, R. M., & Schmidt, H. G. (2004). Detrimental influence of contextual change on spacing effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(4), 796–800.

Vokey, J. R., & Brooks, L. R. (1992). Salience of item knowledge in learning artificial grammars. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(2), 328–344.

Watkins, D. (2001). Correlates of approaches to learning: A cross-cultural meta-analysis. In R. Sternberg & L. Zhang (Eds.)., *Perspective on thinking, learning, and cognitive styles* (pp. 165–195). New Jersey: Erlbaum.

Watkins, O. C., & Watkins, M. J. (1975). Buildup of proactive inhibition as a cue-overload effect. *Journal of Experimental Psychology: Human Learning and Memory*, *1*(4), 442–452.

Weinstein, Y., McDermott, K. B., & Szpunar, K. K. (2011). Testing protects against proactive interference in face–name learning. *Psychonomic Bulletin & Review*, *18*(3), 518–523.

Wells, G. L. (1982). Attribution and reconstructive memory. *Journal of Experimental Social Psychology*, *18*(5), 447–463.

Whittlesea, B. W. A. (1997). *Production, Evaluation, and Preservation of Experiences: Constructive Processing in Remembering and Performance Tasks*. *Psychology of Learning and Motivation, Advances in Research and Theory* (Vol. 37). Elsevier Masson SAS.

Whittlesea, B. W., & Dorken, M. D. (1993). Incidentally, things in general are particularly determined: An episodic-processing account of implicit learning. *Journal of Experimental Psychology: General*, *122*(2), 227–248.

Wilding, J., & Andrews, B. (2006). Life goals, approaches to study and performance in an undergraduate cohort. *British Journal of Educational Psychology*, *76*(1), 171–182.

Wilson, K., & Fowler, J. (2005). Assessing the impact of learning environments on students' approaches to learning: Comparing conventional and action learning designs. *Assessment & Evaluation in Higher Education*, *30*(1), 87-101.

Wissman, K. T., Rawson, K. A., & Pyc, M. A. (2011). The interim test effect: Testing prior material can facilitate the learning of new material. *Psychonomic Bulletin & Review*, *18*(6), 1140–1147.

Wixted, J. T., & Rohrer, D. (1993). Proactive interference and the dynamics of free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(5), 1024–1039.

Woodworth, R. S., & Thorndike, E. L. (1901). The influence of improvement in one mental function upon the efficiency of other functions (I). *Psychological Review*, *8*(3), 247–261.

Wooldridge, C. L., Bugg, J. M., McDaniel, M. A., & Liu, Y. (2014). The testing effect with authentic educational materials: A cautionary note. *Journal of Applied Research in Memory and Cognition*, *3*(3), 214–221.

Yang, C., & Shanks, D. R. (2018). The forward testing effect: Interim testing enhances inductive learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(3), 485–492.

Yang, C., Potts, R., & Shanks, D. R. (2018). Enhancing learning and retrieval of new information: A review of the forward testing effect. *NPJ Science of Learning*, *3*(1), 1–9.

Zeegers P. (2001) Student learning in science: A longitudinal study. *British Journal of Educational Psychology*, 71, 115–132.

Zhang, L. F. (2003). Does the big five predict learning approaches? Personality and Individual Differences, 34(8), 1431–1446.