IMPACT OF MENTAL WORKLOAD ON RATER PEFORMANCE AND BEHAVIOUR

THE IMPACT OF MENTAL WORKLOAD ON RATER PERFORMANCE AND BEHAVIOUR IN THE ASSESSMENT OF CLINICAL COMPETENCE

By WALTER TAVARES, BSc

A Thesis Submitted to the School of Graduate Studies in Partial Fulfillment of the Requirements

for the Degree of:

Doctor of Philosophy

McMaster University HEALTH RESEARCH METHODOLOGY (2014) Hamilton, Ontario, Canada

TITLE: The Impact of Mental Workload on Rater Performance and Behaviour in the Assessment of Clinical Competence. AUTHOR: Walter Tavares BSc (McMaster University).

SUPERVISOR: Dr. Kevin Eva. NUMBER OF PAGES: 145

Abstract

The complexity and broadening of competencies have led to a number of assessment frameworks that advocate for the use of rater judgment in direct observation of clinical performance. The degree to which these assessment processes produce scores that are valid, are therefore vitally dependent on a rater's cognitive ability. A number of theories suggest that many of the cognitive structures needed to complete rating tasks are capacity limited and may therefore become a source of difficulty when rating demands exceed resources. This thesis explores the role of rating demands on the performance and behaviour of raters in the assessment of clinical competence and asks: in what way do rating demands associated with rating clinical performance affect rater performance and behaviour? I hypothesized that as rating demands increase, rating performance declines and raters engage in cognitive avoidance strategies in order to complete the task.

I tested this hypothesis by manipulating intrinsic and extraneous sources of load for raters in the assessment of clinical performance. Results consistently demonstrated that intrinsic load, specifically broadening raters' focus by increasing the number of dimensions to be considered simultaneously, negatively affected indicators of rating quality. However, extraneous demands failed to result in the same effect in 2 of 3 experiments. When we explored the cognitive strategies raters engage under high load conditions we learned of a number of strategies to reduce cognitive work, including idiosyncratically minimizing intrinsic demands (leading to poor inter-rater reliability) and active elimination of sources of extraneous load, explaining both findings. When we induced extraneous load in manner that could not be easily minimized by raters, we also found impairments in rater performance, specifically the provision of feedback.

I conclude that rating demands, whether induced intrinsically or by extraneous sources, impair rater performance affecting both the utility of scores and the opportunity for learner development. Implications for health professions education and future directions are discussed.

Acknowledgments

This journey began with some fortunate opportunities and continued with the support of so many. I need to begin by thanking Dr. Vicki Leblanc who agreed to meet with me when I was considering a future in academia. Thank you Vicki for your mentorship, for the numerous opportunities you have afforded me and for letting me find my own path. To the Wilson Centre community, all the scientists and fellows past and present and especially our lab group, it really is difficult to express how valuable and meaningful an experience, personally and professionally, my time with you has been. I have no greater compliment than to say that I have learned so much from all of you.

To my colleagues and leadership at Centennial College, thank you for taking on more when I needed to take on less. Thank you Renee Kenny, for having the vision to support a fellowship at the Wilson Centre when there really was no mechanism in place to allow that to happen. Thank you Wendy Lund for your friendship, for unselfishly stepping in to allow me time to focus on this work and for arguing with me (you know what I mean). I'm truly grateful.

To my PhD committee, Dr. Kevin Eva, Dr. Geoff Norman and Dr. Shiphra Ginsburg, thank you for your open door, for the fun and stimulating conversations, for the kindness and most of all for your patience. Thank you Kevin for your incredible mentorship; it truly has been an amazing educational, professional and personal journey that I will always be grateful for. I cannot thank you enough for your patience, advice, friendship and commitment.

To my wife Caitlin, this truly would not have been possible without you! We got married, we moved (twice) and had William throughout graduate school. You were constantly a motivation for me and supported me in what truly has been a countless number of ways. Thank you for understanding me, for interrupting me at just the right times, for teaching me how to find balance, for making me laugh and for helping me keep what matters most central in my life. To my son William, thank you for helping me forget everything else around me when we are together. Just thinking of you as I write this brings a smile to my face.

Dedication

I would like to dedicate this dissertation to my mother and father Julieta and Joao Tavares. Both have been my greatest teachers.

Table of Contents

List of Tables and Figures

Chapter 3

Chapter 4

Chapter 5

Preface

I Walter Tavares am responsible for conceptualizing all research projects included in this thesis. I am responsible for all study designs, data collection, data analysis and manuscript writing. I led and prepared any and all research grants, research ethics applications and submissions for publication. I declare that all work is original.

This page is intentionally left blank.

Chapter 1 – Introduction: Rater Based Assessments of Clinical Competence

The provision of safe, effective and patient centered healthcare is valued as fundamental in Canadian society. This serves to drive and shape a number of societal decisions as well as programs of research aimed at informing these goals. For instance, researchers have studied best practices related to medication administration and strategies to optimize team performance in crises while also developing numerous prediction tools and evidence based guidelines to aid clinicians in diagnostic accuracy and treatment plans, respectively. Despite these and many other advances in health care, ensuring the provision of safe and effective care continues to be largely dependent on the individuals who serve in the health professions. This has led researchers to study the human element in the provision of health care extensively. Understanding diagnostic accuracy, for example, or the development of clinical reasoning and technical expertise, have and continue to be studied so that what is discovered can ultimately define best practices and be included in early and continuing educational programs (Eva, 2005; Graber, 2009). In this way, by focusing on clinicians in the health care equation, we can better understand how health care delivered by individuals can be made safer, more effective and patient centered. This came to be recognized as particularly important when a seminal report titled "To Err is Human – Building a Safer Health Care System", by the Institute of Medicine in 1999 described a health care system threatened by the human element of practice (Kohn, et al. 2000). Since then researchers have continued to report that health care providers are indeed responsible for many instances of health care errors (Graber et al. 2002; Croskerry, 2003; Leape and Fromson, 2006), thereby stressing the importance of making accurate decisions about the individuals we allow to deliver health care. This thesis aims to explore the process by which such decisions are made, by focusing specifically on those individuals who are tasked with observing clinician performance and forming judgments regarding their readiness for practice.

After evaluating medical schools in North America, the Flexner Report in 1910 was instrumental in calling attention to the education of physicians and emphasized the need for improved standards, content and process to ensure physicians receive adequate preparation. More recently, the rise of competency based medical education (CBME) places greater emphasis on ensuring that individuals possess meaningful knowledge, skills, and attributes organized around competencies that the profession and public expect of the health professional (Frank, et al. 2010). Biosciences and medical knowledge continue to be important, as they were in the Flexner era, but modern competency frameworks reflect a greater complexity in medical education. For example, educators are additionally emphasizing the value, role and integration of greater social and behavioural sciences, humanities and non-technical skills (Kuper and D'Eon, 2011). The Canadian Medical Association, Medical Council of Canada, Accreditation Council for Graduate Medical Education, Royal College of Physicians of Surgeons of Canada, among other accreditation and licensing or certifying bodies worldwide, have made essential a number of domains of professional practice including medical expert, communicator, collaborator, manager, health advocate, scholar and professional. Both the Flexner era and now the CBME era have gone a long way to shape curriculum and educational practices toward meeting these expectations. However, the community continues to struggle with the need to determine whether or not candidates have indeed achieved what is expected of them within these complicated domains of competence in a manner that is accurate, fair and defensible (Lurie, et al. 2009; Lurie, et al. 2011; Carraccio and Englander, 2013).

The ability to accurately assess clinical competence is as fundamental as the educational content itself (Carraccio and Englander, 2013). The challenges arise in part due to the complexity associated with competency frameworks, but also due to difficulties in understanding what "competence" is, how best to describe it, how it evolves (e.g., meaningful milestones) and how it might be expressed (Kane, 1992; Hodges and Lingard, 2013). As Kane, and more recently, Hodges describe, competence is complex, dynamic, and more than just isolated knowledge, skills and/or judgment, as defined by the profession (Kane, 1992; Hodges and Lingard, 2012). Rather, competence involves the integration of a number of domains of competencies and adapting specific behaviours across a range of possible clinical challenges and encounters (i.e., contexts) (Kane, 1992). Complicating matters further, many domains of competence as currently defined, can be abstract, poorly aligned with the way we naturally think about competence and are difficult to operationalize in ways that promotes or facilitates measurement (Lurie, et al. 2009; Lurie, et al. 2011). For example, in reviewing the literature, Lurie et al., failed to find evidence that "intuitive sounding constructs such as communication skills or professionalism emerge as measureable constructs from the psychometric data" (Lurie, et al. 2011). Much like intelligence, motivation or introversion, competence cannot be measured directly, but rather must be inferred based on behaviours exhibited by individuals in response to a collection of stimuli that are representative of the profession or construct.

Despite these complexities the purpose of summative assessment, in particular, remains to provide an accurate and meaningful indication of a candidate's ability to integrate various competencies and provide a mechanism by which to predict future clinical performance in novel contexts. Toward this goal, Miller provides a useful conceptual framework for structuring assessment strategies (Miller, 1990). In this framework, Miller proposed a four level pyramid where each level represents an increasingly complex standard of performance, informing how each level might be optimally elicited (i.e., the stimuli or task presented to the candidate) and captured (i.e., most appropriate response format). At the base of the pyramid is level one: knows. At this level, stimuli are designed to assess a candidate's declarative or factual knowledge. An effective and efficient stimulus to assess this type of knowledge is often a written or oral exam and the response formats are the answers provided by the candidate. Level two is knows how. At this level stimuli may have more of a clinical focus, involving problem solving or clinical reasoning and are designed to assess a candidate's ability to apply declarative or factual knowledge in a given context. This too can be assessed efficiently and effectively using a written exam but also using more sophisticated tools such as virtual reality / computer based simulators. The answers or actions selected would serve as the response format. Level three is shows how. At this level the stimuli requires action on the part of the candidate. Depending on the competencies targeted or goals of the assessment process, the stimuli may elicit isolated and decontextualized technical skills (e.g., suturing on a part task trainer) or the integration of multiple competencies in a scenario or case based process. Here, as an example, the candidate might be required to elicit necessary information from a patient in an appropriate sequence, to balance other priorities, to perform various technical skills, to work assertively within a team while demonstrating effective professionalism, communication and empathy, and to generate and implement care plans. The stimuli at this level are typically simulation based (e.g., task trainers, mannequins or standardized patients) and the response format is actions performed by the candidates. Finally, at the top of the pyramid is level four: does. At this level candidates are expected to demonstrate the integration of multiple competencies in real clinical contexts with real patients. At this level, the competencies included in the assessment may be

unpredictable but require demonstration of skill in actual clinical practice. This level is, therefore, argued to provide the best opportunity to determine a candidate's ability to integrate various competencies and predict future clinical performance (Epstein and Hundert, 2002).

A number of implications can be drawn from Miller's "pyramid of competence" when considering best practices in assessment. First, any one particular assessment method is not likely to be well suited to assess all levels equally or at all. In the arsenal of assessment strategies and tools that exist, some are more appropriate (i.e., defensible, efficient, effective) than others for a given level or context. The degree to which an assessment strategy is appropriate, becomes, most importantly, a matter of validity (discussed below). This suggests that in order to assess all levels of the pyramid, a number of assessment strategies may be needed. Second, while a given level may be dependent to some extent on the level below, it does not necessarily predict performance at levels above. Adopting Miller's framework suggests then that ideally all decisions regarding clinical competence would involve some assessment in real clinical contexts with real patients (Norcini, 2005; Crossley and Jolly, 2012), or that at the very least, assessment at the "shows how" level be designed and implemented such that it serves as a suitable predictor of performance at the "does" level. Whether assessments take place in simulated or workplace-based settings, assessments at this level are dependent on direct observation of candidates. The degree to which observers/raters are able to effectively complete this task has significant implications for claims of validity associated with the process.

Kane describes validity as the degree of evidence one has favoring the meaningfulness of the inferences or decisions that are made based on the scores generated in an assessment process (Kane, 2006). That is, validity refers to the plausibility, accuracy or appropriateness of a proposed interpretation (e.g., degree of competence) or proposed use (e.g., certification vs. remediation; advancement), and not the test itself. In this framework, when inferring competence using direct observation, anything that influences the scores other that the construct being measured is considered a threat to validity and is referred to broadly as construct irrelevant variance (CIV).

There are a number "cognitive steps" (i.e., inferences) made between a candidate's performance and final trait based decisions regarding clinical competence that may be impacted by CIV. One such step involving raters is referred to as "scoring". Kane describes "scoring" as the process of moving from observation of a candidate's performance in response to a clinical challenge to generating some form of categorical judgment or score describing the candidate's ability (i.e., from an observed performance to an observed score). When assessing at the "shows how" and "does" level, as candidates exhibit behaviours, raters must perform a number of cognitive tasks including actively detecting relevant elements of the performance, evaluating the adequacy with which the performance matches a known standard, assigning appropriate weightings, ignoring irrelevant data, considering contextual influences etc. Raters must then, either as the performance is occurring or immediately after, translate all of this information into some form of judgment. That is, in order to support "scoring" inferences, researchers ought to consider that raters have an active role of attending and processing a significant amount of information and commonly have to do this for multiple domains of competence simultaneously. Given this complexity, raters may introduce a number of idiosyncrasies that can contribute CIV. Until recently, these potential cognitive difficulties have remained underexplored in the assessment of clinical competence, despite being vitally important to validity claims. It is worth noting that "scoring" is only one inference in a chain of inferences described by Kane's validity framework.

Others referred to as "generalization", "extrapolation" and "implication" are equally important and susceptible to various threats. However, I have focused on "scoring" given the direct relationship (described above) to rater performance.

Reliability serves as a quality index in assessment processes by indicating the amount of error associated with the measurement, thereby indicating the ability to use scores to consistently differentiate between candidates (Downing, 2004; Haladyna and Downing, 2004; Eva, 2010). The concept of consistency is important because if scores on a performance assessment reflect the actual ability of the person being observed they should be relatively similar between raters, within raters if tested a second time, across different clinical encounters, etc. The differentiation component suggests the assessment process is designed such that differences between candidates, if present, can be detected. Error is present when an observed score is different than the theoretical "true" score. Error is always present in any assessment process to some extent and can be systematic (i.e., construct irrelevant variance when contributed by raters, as described above) or random (i.e., unpredictable, unidentifiable). Researchers have identified raters as a consistent source of difficulty with various clinical performance assessments that depend on human raters being reported as deficient in terms of inter-rater reliability (Downing, 2005). For example, even after rater training, when faculty were asked to rate the same performances, inter-rater reliability coefficients only ranged between from .34 to .43 (Cook, et al. 2009).

Early efforts to mitigate rater based validity threats have, in large part, attempted to remove the rater from the equation by minimizing human inference and judgment and promoting objectivity and objectification. In this context, objectivity refers to the freedom from judgment and objectification is a set of strategies intended to reduce measurement error (Norman, et al. 1991; Vleuten, et al. 1991). One such strategy involved reducing clinical competence into its component parts and developing itemized dichotomous (i.e., yes / no) checklists or some variant. However, researchers have identified challenges associated with objectification and have concluded mainly that atomization of a competence may lead to trivialization and that objectivity may be illusory (Van Der Vleuten and Schuwirth, 2005). That is, researchers have reported that objectification (1) does not necessarily result in drastic improvements in reliability (Vleuten, et al. 1991), (2) does not necessarily reduce variation between raters (Herbers, et al. 1989; Noel, et al. 1992), (3) that not all that can be measured should be measured (Regehr, MacRae et al. 1998, Hodges, Regehr et al. 1999) and (4) that the simple accumulation of competencies (or checkboxes) does not necessarily equate to clinical competence (Ginsburg, et al. 2010). In other words, rater judgment can be as good as or better than more "objective" assessment strategies, (Regehr, et al. 1998; Hodges, et al. 1999; Eva and Hodges, 2012; Hodges, 2013) but concerns about rater idiosyncrasy remain.

Human judgment is known to be flawed in that raters can have difficulty discriminating beyond two dimensions, can be unacceptably biased, can demonstrate poor intra and inter-rater reliability and are generally considered to be significant sources of error in some contexts (Williams, et al. 2003; Govaerts and Vleuten, 2013). This has led researchers to appropriately question what limits might exist when relying on rater judgment in assessment contexts (Albanese, 2000; Eva and Hodges, 2012). As such, a deeper understanding of rater judgment, rating behaviours or rater cognition is required to further determine how to continue improving upon competence assessment techniques.

Much of the spotlight on rater cognition emerged following research exposing differences between how rating processes were intended and the way in which raters actually behaved when

engaged in the act of assessing clinical performance (Lurie, et al. 2009; Ginsburg, et al. 2010). For instance, researchers in medical education found that rather than emphasizing skills and competencies, as was expected, raters have a tendency to consider a number of non-clinical attributes (e.g., personal qualities, approaches to learning and relational abilities) (Bogo, et al. 2006). Others found that deficiencies or strengths with particular competencies were overlooked depending on other characteristics of the candidate and that perceived competence does not equate to a simple accumulation of competencies or a linear addition of various dimensions (Ginsburg, et al. 2010). Furthermore, despite attempts to standardize or increase objectivity, researchers have come to widely accept that rater based assessments remain associated with subjective influences (Bogo, et al. 2007; Ginsburg, et al. 2010). More recently, researchers have demonstrated that raters idiosyncratically focus on different aspects or dimensions of performance which do not necessarily map neatly onto competency frameworks of standardized rating tools, have difficulty translating behaviours observed to scores on rating tools, see the same thing differently even when assessing the same candidates, apply different performance standards and are subject to numerous biases and external factors (Murphy, et al. 1989; Williams, et al. 2003; Govaerts, et al. 2011; Yeates, et al. 2012).

Collectively, this body of research suggests that raters cannot simply be seen as passive observers or as exchangeable entities, objectively measuring clinical competence. Rather, they must be recognized as unique but important cognitive filters in the process (Landy and Farr 1980). Researchers have done well to describe the performance of raters by identifying the different behaviours and biases, but have not comprehensively explained why such behaviours occur. To date, there is a paucity of research exploring the impact of rating demands or mental or cognitive workload as a causal mechanism for rater idiosyncrasies.

In summary, the provision of safe, effective and patient centered health care is dependent on the individuals who are granted access to the health professions. Allowing candidates to enter the health professions when critical domains of competence have not yet been fully achieved can be costly to patients especially in settings where other safeguards are limited. Current competency based frameworks have made this task more challenging by broadening the domains of competence expected of clinicians while also highlighting that assessment is as fundamental as the education itself. Many of these domains cannot meaningfully be identified through objective means as objectification of clinical assessments has proven limited and rater judgment continues to be valued. Recent research exploring rater behaviours suggests numerous rater idiosyncrasies, but offer limited explanations as to how, why or under what circumstances rater judgment can be used confidently. As rating tasks can be quite complex and may impose high cognitive demands when multiple competencies must be considered this thesis explores the influence of cognitive workload on rater behaviour and performance as a causal mechanism for suboptimal inter-rater reliability. Exploring and understanding causal mechanisms associated with rater behaviour and performance is of theoretical and practical importance and may lead to optimizing assessment strategies so that errors in assessment, such as concluding someone is competent when in fact they are not, can be avoided or at least minimized.

Given the challenges raters appear to experience when rating clinical performance, the complexity with which raters must contend and the limited research contributing causal mechanisms to rater idiosyncrasies, this thesis explores the role of mental workload as it relates to the alignment (or lack thereof) with human cognitive capacity to perceive, attend to and/or process information in these dynamic contexts in real time. A number of cognitive theories

applied to the process of performance appraisal in non-clinical settings suggest raters in clinical settings may be equally at risk of having difficulty managing the cognitive demands associated with the task. Chapter 2, titled "Exploring the Impact of Mental Workload on Rater Based Assessments", applies a focused analysis of literature from a variety of fields (e.g., decision making and judgment, performance appraisal, and various theories in cognitive psychology) informing rater-based assessments. This review exposes mental workload as a potential source of difficulty and suggests that when rating demands exceed cognitive resources, raters are likely to engage in cognitive shortcuts that may yield sub-optimal performance.

To test the resulting hypotheses empirically it was necessary to ensure use of a scale that satisfactorily represented the construct of clinical competence in the context of paramedic training. In chapter 3, titled "Global Rating Scale for the Assessment of Paramedic Clinical Competence" I follow conventions of scale development to develop and critically appraise a global rating scale for the assessment of clinical competence. Scale development has been one of the strategies used to aid raters in the assessment process, but typically does not take into consideration the demands rating tools (or process) impose on raters in practice. This study intentionally replicates common scale development practice to ensure that any struggles identified in later studies do not result from inappropriate construct representation. This study ultimately produced a 7-dimension global rating scale that was used as the rating tool in subsequent studies.

Chapter 4 is titled "The Impact of Rating Demands on Assessments of Clinical Competence". This experimental study uses the conceptual framework from the literature review conducted in chapter 2 and the scale developed in chapter 3 to explore the impact of manipulating rating demands (in a 2x2 factorial design) on rater performance and rating quality. Novice raters were randomly assigned to one of four conditions and asked to rate 3 pre-recorded unscripted clinical encounters illustrating 3 levels of performance (high, medium, low). The number of dimensions participants were asked to rate (7 vs. 2) was manipulated, as was the requirement (or lack thereof) to conduct additional extraneous, but ecologically valid, tasks (attending to patient status and the activity of additional individuals observable on video). Outcome measures included number of dimension relevant behaviours identified, ability to discriminate between levels of performance, and inter-rater reliability. Novice raters were intentionally chosen for this study so that we could test in subsequent studies (chapter 5) whether content knowledge (which typically involves cognitively efficient schemas; mental shortcuts) could serve to mitigate any challenges observed.

Chapter 5 is titled "Selecting and Simplifying: Rater Performance and Behaviour When Considering Multiple Competencies". In this study I used a parallel, mixed methods study design to collect quantitative and qualitative data simultaneously while analyzing the data of each strand concurrently but independently. Faculty raters were recruited as participants and randomly assigned to one of four conditions in a 2x2 factorial design. Factor A was again the number of dimensions (7 vs. 2) to be considered simultaneously and Factor B was the additional requirement (or lack thereof) to rate the performance of standardized actors as well as the candidate being assessed. All participants were asked to observe videotapes of 3 unscripted clinical performances while verbally identifying relevant behaviours and rating the performance using the assigned scale. This study extends the previous study by using more expert raters who are presumed to have more efficient cognitive resources (i.e., schemas) and therefore may be

more resilient to high rating demands. Second, I explored specifically the cognitive strategies raters report to engage under conditions of high demand.

Chapter 6 is titled "Extraneous Demands in Rater Based Assessments of Clinical Performance". In this experimental study I again manipulate rating demands and explore the impact on indicators of rating quality. This time I include an analysis of the feedback provided by raters to learners under different load conditions. Expert raters were randomly assigned to either the role of "passive rater" or "immersed rater" in a simulation based, performance-based exam. Participants assigned to the passive condition were tasked simply with observing and rating trainee performances using a 7 dimension global rating scale (GRS). Participants assigned to the immersed condition were similarly tasked with observing and rating trainee performances but were "immersed" in the case meaning they were required to additionally play the role of a standardized actor and guide the direction of the simulation as necessary. Immediately following the case, all participants were required to rate the performance and document feedback they would give to support trainee development. Outcome measures included reliability and the amount and type of feedback given.

Finally, in Chapter 7 I review and summarize these findings, draw conclusions regarding the impact of rating demands on rater based assessments of clinical competence, discuss implications for health professions education and outline future directions aimed at further advancing the field.

Ethical Considerations

In each of the experimental studies described above Research Ethics Board (REB) approval was received. All participants provided informed consent, were offered the opportunity to withdraw at any point and we assured of anonymity / confidentiality throughout. All data continues to be stored and secured until as per our REB requirements. We did not experience any ethical issues or concerns from any of the REB boards we sought approval from for any of the work completed for this thesis.

Chapter 2 – Literature Review: "Exploring the Impact of Mental Workload on Rater Based Assessments"

Published: Tavares, W., Eva, K.W. (2012). Exploring the Impact of Mental Workload on Rater-Based Assessments. *Advances in Health Sciences Education*, 18(2): 291-303.

Abstract

When appraising the performance of others, assessors must acquire relevant information and process it in a meaningful way in order to translate it effectively into ratings, comments, or judgments about how well the performance meets appropriate standards. Rater-based assessment strategies in health professional education, including scale and faculty development strategies aimed at improving them have generally been implemented with limited consideration of human cognitive and perceptual limitations. However, the extent to which the task assigned to raters aligns with their cognitive and perceptual capacities will determine the extent to which reliance on human judgment threatens assessment quality. It is well recognized in decision-making research that, as the amount of information to be processed increases, judges may engage mental shortcuts through the application of schemas, heuristics, or the adoption of solutions that satisfy rather than optimize the judge's needs. Further, these shortcuts may fundamentally limit/bias the information perceived or processed. Thinking of the challenges inherent in rater-based assessments in an analogous way may yield novel insights regarding the limits of rater-based assessment and may point to greater understanding of ways in which raters can be supported to facilitate sound judgment. This paper presents an initial exploration of various cognitive and perceptual limitations associated with rater-based assessment tasks. We highlight how the inherent cognitive architecture of raters might beneficially be taken into account when designing rater-based assessment protocols.

Introduction

Assessment of clinical performance remains both a priority and a major challenge in health professional education (Huwendiek et al., 2010). High quality assessments facilitate appropriate inferences regarding an individual's level of competence or performance and provide an opportunity for meaningful feedback. The quality of assessment has been recognized as a key element in promoting the use of competency-based outcomes, facilitating fairness, building credibility and defensibility for licensing bodies, and engendering trust from the end user (e.g., public, hospital, clinic, and the practitioners themselves). However, challenges persist and researchers continue to investigate strategies to improve assessment protocols.

One such challenge derives from the fact that a variety of domains of competence are sufficiently abstract as to require the judgment of evaluators. Challenges associated with attempts to objectify complex and dynamic clinical performances have led the assessment community to recognize and value expert rater judgment as evidenced by the adoption and use of global ratings in assessment protocols (Hodges et al., 1999). However, rater based assessments are fallible and subject to significant difficulties (Morgeson and Campion, 1997) such as observational inaccuracy, common rating biases (e.g., halo error, end aversion, and positive skew), failure to identify deficiencies and difficulty discriminating between dimensions (Herbers et al., 1989; Thompson et al., 1990; Kalet et al., 1992; Noel et al., 1992; Haber and Avins, 1994; LaMantia et al., 1999; Williams, et al., 2003; Lurie, et al., 2009;). Generalizability studies consistently identify raters as a significant source of error (Downing, 2005); Margolis, et al. 2006; Cook, et al. 2010). Various explanations for these rater difficulties have included differences in the concepts raters possess, failure to follow instructions, lack of training, unfamiliarity with the scale and lack of practice. Attempts to address these rater difficulties have traditionally taken the form of scale development and rater training. However, neither strategy (alone or in combination) has resolved the persistent challenges. This has sparked debate over the most suitable approach to assessment and identified a need for novel investigations (Govaerts et al., 2007).

One under-explored line of inquiry aimed at improving rater-based assessments involves the study of mental workload and its influence on rater cognition. Mental workload refers to the cognitive effort expended during a particular task. To function effectively, exogenous demands imposed by the task must be aligned with the supply of endogenous attentional and processing resources needed to manage it. In this way, mental workload is inextricably linked with cognitive functioning (i.e., the mental processes or strategies associated with performing a given task).

Evidence suggesting that raters have difficulty with the mental workload demands imposed by rating tasks becomes less surprising when one considers the complexity inherent in rating clinical performances. In fact, a series of cognitive models in the performance appraisal literature have evolved to demonstrate just how complex a process rating a performance can be as well as indicating how mental workload and rater cognition become intertwined (Borman, 1978; Cooper, 1981; DeNisi et al., 1984; Feldman, 1981; Ilgen and Feldman, 1983). Applied to the assessment of clinical performance, a rater is expected to carefully attend to behaviours in order to actively detect and select relevant elements of the performance to process, assimilate and categorize in

working memory. The rater may then need to retrieve integrated information regarding the standards of the profession and the norms for the candidate's level of training from long term memory to translate their observations into evaluations of the adequacy with which the skills are performed while assigning appropriate weightings to each aspect of performance. For most rater-based assessments, raters have to do this not for one isolated complex dimension, but for many simultaneously. For example, the commonly used mini-CEX includes 6 dimensions: interviewing, physical examination, humanistic qualities / professionalism, clinical judgment, counseling and organization/efficiency (Norcini et al., 2003).

Such multiplicities of focus suggest that mental workload demands and complex rater cognitive processes play an important role in rater-based assessment and that all information must pass through a "cognitive filter represented by the rater" (Landy and Farr, 1980). Raters must therefore be viewed as seekers of information and information processors who bring to the rating tasks cognitive properties that are independent of the performance elements presented to them and vulnerable to inherent cognitive limitations (Ilgen et al., 1993; Landy and Farr, 1980). Precisely how mental workload and these thought processes influence raters and what can be done to support them is less well understood.

This paper attempts to further understand these issues by presenting an initial exploration of various cognitive and perceptual processes associated with rater-based assessment tasks. While others began this exploration (e.g., Williams, Klamen, and McGaghie, 2003) our goal is to expand the discussion by focusing specifically on the link between mental workload, inherent cognitive architecture and rater cognitive operations related to rating tasks. Thinking of the challenge of rater based assessments in this way can lead to novel insights regarding the limits of rater based assessments and a need to consider rater based errors that extend beyond social or environmental influences. Taking into account the mental workload demands imposed on raters during rating tasks and the inherent cognitive capacity of raters can expose cognitive and perceptual capacity limitations as a critical element in assessment and may therefore, provide insights into how educators and researchers can continue to address the challenges associated with assessment of clinical performance.

With this in mind, we structure this paper by first reviewing a set of theoretical frameworks from cognitive psychology that are applicable to mental workload and its influence on information acquisition and processing. Given that many programs of assessment and scale development strategies are limited in their consideration of these issues, we focus specifically on perceptual and processing loads that may create difficulties noticing, attending to and processing of relevant behaviours. Because the question to be addressed is broad and requires consideration from multiple angles of a diverse literature, systematic review techniques are ill suited to our research question (Eva, 2008). Instead, we conducted a critical review aimed at synthesizing research from a variety of perspectives. While this review cannot be summarized comprehensively in the format of a journal article we have attempted to include representative samples of theoretical positions and empirical findings that appear to be particularly relevant to the issue of rater-based assessment. Based on the findings we then derive implications for health professional educators and health professional education researchers while recognizing that systematic testing of these ideas needs to be conducted in that context.

Attention

In order for stimuli to be consciously perceived and processed at least to the point where they can be reported, the stimuli must be attended to (Mack and Rock, 1998). Attention has been referred to as selectivity of processing and explained as "taking possession of the mind in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought...it implies withdrawal from some things in order to deal effectively with others" (James, 1890; Pashler, 1998; Levitin, 2002). Attention as a limited resource plays two critical roles. As a selective agent, it dictates what information reaches an information processing stage by choosing and constraining the information that will be perceived (Pérez Moreno et al., 2011; Marois and Ivanoff, 2005; Miller, 1956). As a task management agent, it facilitates ongoing processing (Pérez Moreno et al., 2011) by ultimately providing working memory access to and control over the information available, but also constrains what tasks can be performed concurrently. The implications of this fundamental aspect of cognition are extensive.

What and how much is attended to, and by extension what remains undetected, is influenced by the amount of perceptual load placed on the individual. Perceptual load theory states that under high load conditions attentional resources are drained and stimuli, especially non-task relevant stimuli, (e.g., aspects of performance distinct from those attended to) are likely to be missed (Lavie, 1995). That is, focusing attention on a stimulus can impair perception of other stimuli especially when the task attended to involve a high level of perceptual load and consumes all available capacity. This phenomenon is known as 'inattentional blindness' (Mack, 2003; Pérez Moreno et al., 2011; Simons, 2000) and suggests that the gains made by work-based assessment in terms of embedding assessments in an authentic environment may be offset by the considerable downsides of increased perceptual load.

Attention and Rater-Based Assessment

Evaluating mental workload as it relates to acquisition and processing limits can provide insights into the way in which judgments are ultimately made about clinical performance. In reviewing the literature, Holmboe provides evidence to suggest that the quality of faculty *observation skills* (i.e., information acquisition) is lacking (Holmboe, 2004). In doing so, he argues that training is required, but it is also possible that fundamental limits in perceptual capacity may help explain some of the deficiencies. As a result, modification of the assessment protocols themselves may be required. Holmboe (2004) cites a study by Herbers et al., (1989) designed to evaluate how well raters observe clinical performances. This study involved having faculty rate clinical performances in a simulated setting and found that raters varied significantly in the acuity of their observations (Herbers et al., 1989). Further, and relevant to our thesis, it was shown that *what* faculty attended to also varied significantly. Herbers (1989) implies the results of this study (i.e., observational accuracy) were worse than expected because the protocol required faculty to observe many specific elements of performance, suggesting that the demands were greater, relative to rating a single aspect of competence.

Reinterpreting these findings through the lens of perceptual load theory would suggest that attending to certain elements of performance drains attentional capacity resulting in elements of performance that, for all intents and purposes may be self-evident, are simply idiosyncratically not "seen". In other words, information that otherwise may have contributed to the rating is lost and, as a result, significant variation between raters can be expected. Consistent with this suggestion, is a study conducted by Lamantia et al., (1999) that evaluated inter-observer

variability among faculty evaluating residents and found significant variation in the observation of clinical performances regardless of whether raters used checklists or global rating scales (LaMantia et al., 1999). Lamantia et al., (1999) concluded that the differences between observers were not a function of the scale or the performance domains, but rather, resulted from "fundamental discrepancies among evaluators in the capacity to observe specific behaviours" (LaMantia et al., 1999).

In sum, the perceptual load imposed on raters may easily exceed their attentional capacity given the amount of information present in even a brief encounter, the rate at which the information must be processed, the potential for interactions between elements of performance, and the demands imposed by typical assessment protocols to translate one's observations into multi-dimensional ratings. The impact of inherent cognitive limitations will inevitably be idiosyncratic and, hence, will introduce construct irrelevant variance even if each rater, while variable in their impressions, is correct in their observations.

Information Processing

We must also concern ourselves with how information that makes it past perceptual bottlenecks is mentally processed by the rater. Information processing theory describes behaviour in terms of the interaction between an individual's cognitive system, the task at hand, and the problem space (i.e., the individual's internal representation of the task) (Payne, 1980). This interaction between the task and internal representations held in long-term memory is believed to take place within a limited capacity working memory system (Baddeley, 2007). A central executive coordinates the interactions, controlling the selection and integration of information. Analytic procedures (e.g., mental arithmetic), the transformation of information (e.g., re-working one's sentences when writing a paper), and the identification of links between the current task and previous experiences stored in long term memory (i.e., analogical transfer) all take place within this system (Lord and Maher, 1990) and distinguish information processing from perceptual tasks. The load or processing demands imposed on this system are influenced by the number of processes occurring simultaneously and their interactions, the time required for each, and the amount of work needed to process the information.

Both working memory and the central executive are limited in how many operations they can execute at one time. Mental workload can overwhelm processing capacity in a dynamic environment, resulting in constrained processing ability (Wickens and Carswell, 1988). Cognitive Load Theory (CLT) has been put forward to explain limitations associated with information processing and their impact on performance. It is based on a human architecture consisting of an unlimited long term memory and a working memory system limited in terms of how long and to what extent it can actively work with a certain amount of information (Lord and Maher, 1990; Van Merriënboer and Sweller, 2010). Generally, in educational contexts, the goal is to reduce extraneous load (i.e., mental activities not directly related to learning) to free cognitive resources for intrinsic load (i.e., to be learned information) and germane load (i.e., mental effort directly related to the learning task). CLT is most commonly used to identify ways to align exogenous demands with endogenous resources for optimal learning (Mayer, 2010). This can be done externally through instructional design such as presenting a particular type, format or amount of information, or internally by the individual with varying degrees of effectiveness (Bannert, 2002).

When strategies to minimize or reduce mental load are unavailable or ineffective however, or as the information becomes increasingly complex, the potential for overload increases and can lead to performance impairments, avoidance and/or simplifying strategies (Lord and Maher, 1990; Morgeson and Campion, 1997; Beckmann, 2010). For instance, cognitive processes associated with situation awareness in aviation are similar to that of rater cognition in that both require similar levels of perception and comprehension of dynamic stimuli presented in live time (Tsang and Vidulich, 2002). When cockpit designs and pilot tasks were manipulated to influence mental workload, a negative correlation between mental workload and situation awareness was found (Alexander, 2000). Similarly, when performance in laparoscopic skills was measured along with mental workload, as mental workload scores increased, *performance* impairments (e.g., inadvertent injuries to adjacent structures) were observed (Yurko et al., 2010).

In another set of studies it was found that individuals tend to migrate towards lower demand conditions. When participants were asked to complete one of two simultaneously presented arithmetic tasks, one involving a carrying operation, which increases computational complexity and the other without, participants *avoided* the high demand condition the majority of the time (Kool et al., 2010). Similarly, in a series of laboratory-based experiments where the load associated with decisions was manipulated, decision makers had a tendency to avoid tasks carrying higher cognitive demands (Botvinick and Rosen, 2009). As a means to reduce effort associated with cognitively demanding tasks and circumvent the limitations of working memory individuals may engage heuristics (i.e., mental shortcuts) as an effort-reduction strategy (Shah and Oppenheimer, 2008), and/or schemas (i.e., automatic and stable prior knowledge held in long term memory) (Lord and Maher, 1990). While the use of schemas and heuristics are generally considered useful in that they can lead to satisfactory and efficient outcomes, they may also contribute to judgment error (Tversky and Kahneman, 1974). In contrast to research in medical decision-making and clinical reasoning, there is a paucity of research applying these concepts to the act of rating clinical performances. Next we directly consider how the limits of our cognitive system, described above, might apply to that context.

Information Processing and Rater-Based Assessments

Considering rater-based assessment through the lens of CLT, intrinsic load would involve the consideration of multiple elements of performance and their interactivity, extraneous load would include any competing mental activities (e.g., ignoring irrelevant elements of the individual being assessed, controlling a simulator, role playing a patient scenario, monitoring patient safety, or managing a busy clinic) and germane load, similar to learning, would be the mental activity or effort associated with the rating task. Aligning mental workload with mental resources is challenging given the inherent features associated with typical rating tasks. While some forms of load can be easily modified (e.g., using a simulation technologist to control a simulator instead of having the assessor manage both the case and simulator) others (e.g., the tendency/need to consider patient care while also attending to the trainee) are more challenging to manipulate leaving us with the conundrum of how to help the rater consider multiple competencies to the exclusion of construct irrelevant aspects of performance.

Even when sources of extraneous load are minimized, rating task complexity can overwhelm processing ability and threaten the alignment principle. Complexity is directly related to the number of distinct processes that are to be executed in completing the rating in addition to the number of distinct information cues that must be processed in order to work through the rating

process. It follows that mental workload (i.e., processing demands) would be higher when multiple elements of performance for multiple dimensions must be processed simultaneously and that, as a result, ratings may ultimately be affected (Kogan et al., 2009). When Melcher et al., (2010) asked raters to rate either 4 or 5 candidates for managerial positions using the same 4 dimensions in leaderless group tasks, they found significant impairments of assessors' rating quality when higher demands (rating 5 candidates) were compared to lower demands (rating 4 candidates) (Melchers et al., 2010). This is a dramatic finding given that the manipulation was so minimal. These researchers concluded that discriminant validity between dimensions is better in lower demand tasks.

It appears then that the amount of load induced by a rating task may be manipulated with predictable outcomes. As another example, although load per se was not explicitly discussed in a study comparing novice and expert raters, researchers were able to manipulate rating task complexity by increasing stimuli length from 6 to 18 minutes, and by increasing the complexity of student behaviours raters were expected to evaluate (Govaerts et al., 2011). These differences led to differences in rater cognition. For the low complexity tasks, performance was similar between expert and novice raters (the former group presumably having greater free cognitive capacity due to the automation that occurs with expertise), but differences were observed for the more complex behaviour pattern task (Govaerts et al., 2011). This suggests that not only will complexity influence mental workload, thereby potentially shedding further light on the reason that multiple brief exposures to individual performances have been found in other contexts to yield better psychometric outcomes than more elaborated performances (Eva et al., 2004), but also provides an indication that one's capacity to deal with such load changes with expertise. Information processing theory would suggest that other factors, such as number of performance elements exhibited by the candidate, how the information was presented (i.e., order, organization, completeness, quality), speed at which the information was presented, and associated processing requirements (e.g., patient complexity) may also contribute to differences in complexity. Better understanding the extent to which these factors influence rating quality may yield better assessment by reducing the extraneous load with which examiners are forced to grapple.

When mental workload associated with ratings is elevated and alignment between demands and resources cannot be improved through external strategies, raters are left to manage load through internal strategies. As described earlier, under conditions of high complexity, individuals are likely to avoid cognitively complex tasks. If this is true in the context of rating clinical performances, raters may engage cognitive strategies such as serial processing or degraded concurrent processing (Wickens and Carswell, 1988). Serial processing would have the rater processing some dimensions for example, while neglecting or delaying others under conditions of high mental workload. Degraded concurrent processing would have the rater processing multiple dimensions at a lower level of accuracy than if processed in isolation.

Engaging schemas and heuristics as an internal strategy likely improves efficiency and may result in highly effective ratings. When applied to limit the amount of information with which the rater is required to work, they can also be problematic. Schemas are based on individual experiences, and application of these personal constructs can lead to significant variation between raters that otherwise may not have been present under lower load conditions. Similarly, the application of heuristics may result in raters examining fewer cues, resulting in the

integration of less information and differences in the foundation on which judgments are formed across raters (Shah and Oppenheimer, 2008). Both may contribute error and/or lead to reduced discrimination.

In summary, variation between raters is likely to exist in the elements that are perceived and/or processed due to perceptual and/or processing demands associated with a rating task. Attentional capacity limits make observing the assortment of behaviours difficult and can lead to variations in rating. Attempts by the rater to avoid complexity internally may result in the use of mental short cuts and contribute further to differences between raters. This variation between raters may indicate a beneficial diversity of perspective, but may also point to ways in which rating tasks can seem non-credible. Appreciating various sources of mental workload, factors contributing to complexity and the implications of each, can provide a framework by which to target strategies for improving assessments. That said, unlike opportunities that can exist for reducing load in other fields (e.g., multimedia learning tasks, air traffic controllers), in some contexts (e.g., in-training assessments) there may be less room for manipulation. Where authenticity is valued, understanding the effects of overwhelming raters and the inherent limits they place on assessment can be beneficial.

Discussion and Future Directions

When difficulties arise and rater based assessments result in poor rating quality two main categories of explanations are provided; those attributed to scales and those attributed to raters. For example, in reviewing the literature, Lurie et al. (2009) found that socially constructed competencies do not lend themselves well to measurement (Lurie et al., 2009). They suggest that a link between that which can be measured and the more general overarching competencies listed by the Accreditation Council for Graduate Medical Education (ACGME) is needed and that, if the social construction of competence was adequately understood and appreciated, measurement strategies may be improved (Lurie et al., 2009). In contrast, in two separate studies, the American Board of Internal Medicine noted challenges in raters' ability to discriminate between dimensions and attributed their results not to the measurement tools, but rather to common rater errors such as halo and leniency (Haber and Avins, 1994; Thompson et al., 1990). As a result, Green and Holmboe (2010) argued that the issue is not the assessment instruments, but rather, is the faculty who use them. Rather than trying to develop the perfect scale, they suggested rater training as a solution to the improper use of scales by inexperienced faculty (Green and Holmboe, 2010).

In an effort to contribute to the debate, we suggest that a middle ground might be found by expanding the discussion regarding error associated with rater based assessment to involve the perceptual and processing loads imposed by rating tasks and the inherent cognitive limitations that may exist in response to those loads. Researchers have argued a dissonance exists between the way competencies are represented on rating forms and the mental frameworks raters actually use (Bogo et al., 2004; Ginsburg et al., 2010; Regehr et al., 2007), that raters engage schemas (Govaerts et al., 2007), that ratings could improve through memory aids (Williams et al., 2003) and that aligning performance assesments with constructs held by raters enhances reliability and validity (Crossley, et al. 2011). We suggest extending this type of research to exploring the alignment between cognitive resources and demands imposed by a rating task.

We have emphasized two main sources of variability in rater performance, specifically the process of information acquisition and information processing as necessary precursors to translating one's observations into a judgment. Our review of the literature suggests that in some forms of clinical performance assessment, the demands imposed on raters may overload cognitive capacity and thus lead to selection and detection variation during information acquisition with subsequent processing strategies applied in attempts (often implicit) to reduce mental workload. As a result, we do not suggest abandoning either scale development or rater training research. Rather, we argue that studying the role of cognitive capacity at information acquisition and processing stages might help advance both efforts by further elucidating the grounds on which judgments are ultimately made. We suggest first that researchers evaluate the mental workload associated with rater based assessments in order to further understand the context in which raters are expected to function, then to carefully reduce mental workload where and when possible.

Evaluate Load

Researchers should investigate the perceptual and processing loads associated with various rating tasks in natural and context specific settings. The most common ways of doing so include (i) subjective ratings, (ii) primary task performance and (iii) secondary task performance. Subjective rating scales (i.e., self-report measures) require individuals to report on the mental effort required to perform a particular task (Paas, et al. 2003). Researchers have found that individuals are capable of providing such ratings (Gopher and Braune 1984, Paas, et al. 2003). Primary task performance measurement considers changes in performance on the actual task (e.g., rating quality, inter-rater agreement, discrimination between dimensions) as a function of changes in task demands. This approach likely offers the greatest face validity as it most directly links cognitive capacity with rater-based assessment strategies. For example, one could manipulate load by altering the number of dimensions to be rated or by including additional tasks (e.g., controlling a simulator). Disadvantages of this strategy include the omnipresent challenge of rater-based assessments being required precisely because there are no gold standard objective measures that can define with certainty the accuracy of the decisions reached, thereby causing one to rely on proxy measures like inter-rater reliability. Finally, secondary task performance measurements are those collected on tasks performed concurrently with the primary rating task. For example, the more unused capacity one has when completing a rating task the better one should perform on a different task like monitoring a patient's heart rate or other vitals (Paas, Tuovinen et al. 2003, Tsang and Vidulich 2006). Measurement of mental workload has been applied extensively in the aviation industry and is beginning to be used in the health professions (Young et al., 2008; Davis et al., 2009; Byrne et al., 2010; Yurko et al., 2010; Zheng et al., 2010; Schulz et al., 2011). A comprehensive treatment of strategies for mental workload measurement is provided by Tsang and Vidulich (2006) and Paas et al. (2003).

Programs of assessment, scale development and rater training have for the most part paid little attention to these cognitive issues, as issues related to reliability (e.g., ensuring multiple observations) and validity (e.g., ensuring the construct is adequately represented) have dominated the assessment literature. Both may stand to be improved by research into questions such as: What circumstances lead to excessive mental workload and how does that influence the rating task? What, if anything, can be done to minimize the detrimental impact of load? What are raters able to attend to under different circumstances (e.g., when patient safety demands are concurrent with rating demands)? What can be done to help assessment designers anticipate and

measure the amount of load and its impact on assessment? Answering these questions can help identify sources of construct irrelevant variance and move the field toward improved rating quality.

Reduce Load

Effective rating can only occur if, like effective instructional designs, the rating procedure aligns with the cognitive capacity (in terms of acquisition and processing) of the rater. Until a strategy is designed to enable load to be evaluated in each context, it might be beneficial to incorporate strategies that provide a means by which to manage load externally while being careful not to reduce ratings to a collection of simple isolated tasks that breakdown the value of rater judgment. This is especially true in the setting of work-based assessments where environments are variable and unpredictable with many competing demands. For example, separating 'clinician' and 'rater' roles in work-based assessments, allowing the faculty member to focus on the candidate's performance, would reduce extraneous load (mental activities not directly related to the rating task) freeing resources for intrinsic load (tasks directly relevant to the rating task) and germane load (mental effort dedicated to the rating task). Whether or not a trained clinician can avoid focusing on the clinical details in a patient scenario remains to be tested.

Another strategy may include reducing complexity where possible. For example, test administrators may allow the candidate to perform in an authentic clinical encounter (where multiple competencies are likely demonstrated simply as a function of the interaction), but require raters simply to evaluate a subset of all possible dimensions to lessen the chance of exceeding attentional or processing capacity. For example, one rater may be tasked with evaluating communication skills and professionalism, while another assesses physical exam skills. While feasibility issues definitely need to be grappled with, the intention would be to take cognitive load into consideration to reduce demands that may be problematic. Hinsz et al., (1997) introduced a similar concept for learners, which is that a collaborative group of learners could be considered an information processing system consisting of multiple, limited working memories that can create a collective working space. In this way, the cognitive load associated with carrying out a rating task could be distributed across multiple collaborating memories or attention systems, reducing the risk of overloading each member and creating a larger reservoir of cognitive resources (Hinsz et., 1997). Such strategies have been found effective in the context of improving assessment of autobiographical submissions in an admissions context (Dore et al., 2006). While necessarily speculative at this point as direct evidence has not been collected in the broader world of assessment relevant to medical education, researchers in other fields investigating similar concerns have demonstrated that reducing load may create strategies for improved rater performance (Gaugler and Thornton, 1989; Melchers et al., 2010).

Conclusion

Our review of the cognitive tasks associated with evaluating candidates and the evidence indicating considerable effects of high cognitive and perceptual demands suggest that understanding raters' cognitive capacity in an assessment context could have significant implications for working towards improving rater-based assessments. The multifaceted nature of the construct of clinical performance, the reliance on rater judgment and the unique and dynamic settings in which clinical assessments are expected to take place, warrant a field specific study of the alignment between mental workload and cognitive capacity in medical education (see

Crossley and Jolly, 2012 for related arguments). The research described here has both practical and theoretical implications for medical education contexts. On the practical side, we have argued for a new focus that involves evaluating mental workload associated with rater-based assessments. This would allow researchers to identify where and when construct irrelevant variance attributed to perceptual and cognitive load may exist and where threats to reliability emerge. By aligning rating tasks with perceptual and processing capacities, raters (or assessment programs) may demonstrate improved discrimination between dimensions and/or candidates and improved inter-rater reliability. Further, the field might gain a better sense of what variability should be treated as "error" as opposed to meaningful differences in the variability that will inevitably be observed. Load measurements should inform assessment protocols, scale development and research involving rater training.

A second practical implication involves providing the impetus for studying how to effectively reduce load externally. We have provided examples of such strategies and encourage researchers to explore these and other strategies in their own contexts to maintain the integrity of their assessment process, avoid overly reductionist strategies, and optimize levels of intrinsic and germane load while being mindful of perceptual load. On the theoretical side, by extending perceptual load theory, cognitive load theory and information processing theory to clinical performance ratings we have aimed to broaden perspectives on strategies for improving rating quality. These issues cast light on an understudied aspect of rater-based assessment that may be particularly helpful in moving the field forward. With any rating task we can be certain a number of mental processes are involved in the rating of clinical performances. Understanding the vulnerabilities of these processes and capitalizing where possible on their strengths and minimizing weaknesses can provide an important key to effective and high quality rater-based assessments.

References:

Alexander, A. L., Nygren, T. E., & Vidulich, M. A. (2000). Examining the relationship between mental workload and situation awareness in a simulated air combat task: (Tech. Rep. No. AFRL-HE-WP-TR-2000-0094). *Wright-Patterson Air Force Base, OH: Air Force Research Laboratory*.

Baddeley, A. D. (2007). *Working memory, thought and action*. Oxford: Oxford University Press.

Bannert, M. (2002). Managing cognitive load-recent trends in cognitive load theory. *Learning and Instruction, 12*(1), 139-146.

Beckmann, J. F. (2010). Taming a beast of burden-On some issues with the conceptualisation and operationalisation of cognitive load. *Learning and Instruction, 20*(3), 250-264.

Bogo, M., Regehr, C., Power, R., Hughes, J., Woodford, M., & Regehr, G. (2004). Toward new approaches for evaluating student field performance: Tapping the implicit criteria used by experienced field instructors. *Journal of Social Work Education, 40*(3), 417-428.

Borman, W. C. (1978). Exploring upper limits of reliability and validity in job performance ratings. *Journal of Applied Psychology, 63*(2), 135-144.

Botvinick, M. M., & Rosen, Z. B. (2009). Anticipation of cognitive demand during decision-making. *Psychological Research, 73*(6), 835-842.

Byrne, A., Oliver, M., Bodger, O., Barnett, W., Williams, D., Jones, H., et al. (2010). Novel method of measuring the mental workload of anaesthetists during clinical practice. *British Journal of Anaesthesia, 105*(6), 767-771.

Cook, D. A., Beckman, T. J., Mandrekar, J. N., & Pankratz, V. S. (2010). Internal structure of mini-CEX scores for internal medicine residents: factor analysis and generalizability. *Advances in Health Sciences Education, 15*(5), 633-645.

Cooper, W. (1981). Ubiquitous halo. *Psychological Bulletin, 90*(2), 218-244.

Crossley, J., Johnson, G., Booth, J., & Wade, W. (2011). Good questions, good answers: construct alignment improves the performance of workplace based assessment scales. *Medical Education, 45*(6), 560-569.

Crossley, J., Jolly, B. (2012). Making sense of work-based assessment: Ask the right questions, in the right way, about the right things, of the right people. *Medical Education, 46*(1), 28-37.

Davis, D., Oliver, M., & Byrne, A. (2009). A novel method of measuring the mental workload of anaesthetists during simulated practice. *British Journal of Anaesthesia, 103*(5), 665-669.

DeNisi, A., Cafferty, T., & Meglino, B. (1984). A cognitive view of the performance appraisal process: A model and research propositions. *Organizational Behaviour and Human Performance, 33*(3), 360-396.

Dore, K. L., Hanson, M., Reiter, H. I., Blanchard, M., Deeth, K., & Eva, K. W. (2006). Medical school admissions: enhancing the reliability and validity of an autobiographical screening tool. *Academic Medicine, 81*(10), S70-73.

Downing, S. (2005). Threats to the validity of clinical teaching assessments: What about rater error? *Medical Education, 39*(4), 353-355.

Eva, K. W. (2008). On the limits of systematicity. *Medical Education, 42*(9), 852-853.

Eva, K. W., Rosenfeld, J., Reiter, H. I., & Norman, G. R. (2004). An admissions OSCE: the multiple mini interview. *Medical Education, 38*(3), 314-326.

Feldman, J. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. *Journal of Applied Psychology, 66*(2), 127-148.

Gaugler, B., & Thornton, G. (1989). Number of assessment center dimensions as a determinant of assessor accuracy. *Journal of Applied Psychology, 74*(4), 611-618

Ginsburg, S., McIlroy, J., Oulanova, O., Eva, K., & Regehr, G. (2010). Toward Authentic Clinical Evaluation: Pitfalls in the Pursuit of Competency. *Academic Medicine, 85*(5), 780-786.

Gopher, D., & Braune, R. (1984). On the psychophysics of workload: Why bother with subjective measures? Human Factors, 26(5), 519-532.

Govaerts, M., van der Vleuten, C., Schuwirth, L., & Muijtjens, A. (2007). Broadening perspectives on clinical performance assessment: rethinking the nature of in-training assessment. *Advances in Health Sciences Education, 12*(2), 239-260.

Govaerts, M. J. B., Schuwirth, L. W. T., van der Vleuten, C. P. M., & Muijtjens, A. M. M. Workplace-based assessment: effects of rater expertise. *Advances in Health Sciences Education, 16*(2), 151-165.

Green, M. L., & Holmboe, E. (2010). Perspective: The ACGME Toolbox: Half Empty or Half Full? *Academic Medicine, 85*(5), 787-790.

Haber, R., & Avins, A. (1994). Do ratings on the American Board of Internal Medicine Resident Evaluation Form detect differences in clinical competence? *Journal of General Internal Medicine, 9*(3), 140-145.

Herbers, J., Noel, G., Cooper, G., Harvey, J., Pangaro, L., & Weaver, M. (1989). How accurate are faculty evaluations of clinical competence? *Journal of General Internal Medicine, 4*(3), 202-208.

Hinsz, V. B., Tindale, R. S., & Vollrath, D. A. (1997). The emerging conceptualization of groups as information processors. *Psychological Bulletin, 121*(1), 43-64.

Hodges, B., Regehr, G., McNaughton, N., Tiberius, R., & Hanson, M. (1999). OSCE checklists do not capture increasing levels of expertise. *Academic Medicine, 74*(10), 1129-1133.

Holmboe, E. (2004). Faculty and the observation of trainees' clinical skills: problems and opportunities. *Academic Medicine, 79*(1), 16-22.

Huwendiek, S., Mennin, S., Dern, P., Friedman Ben-David, M., Van Der Fleuten, C., Tonshoff, B., Nikendei, C. (2010). Expertise, needs and challanges of medical educators: Results of an international web survey. *Medical Teacher, 32*(11), 912-918.

Ilgen, D., Barnes-Farrell, J., & McKellin, D. (1993). Performance appraisal process research in the 1980s: what has it contributed to appraisals in use? *Organizational Behaviour and Human Decision Processes, 54*(3), 321-321.

Ilgen, D., & Feldman, J. (1983). Performance appraisal: A process focus. *Research in Organizational Behaviour*. (pp. 141-197). Greenwich, CT. JAI Press.

James W. (1890). The Principles of Psychology. New York: Holt.

Kalet, A., Earp, J., & Kowlowitz, V. (1992). How well do faculty evaluate the interviewing skills of medical students? *Journal of General Internal Medicine, 7*(5), 499-505.

Kogan, J. R., Holmboe, E. S., & Hauer, K. E. (2009). Tools for direct observation and assessment of clinical skills of medical trainees: a systematic review. *JAMA, 302*(12), 1316-1326.

Kool, W., McGuire, J. T., Rosen, Z. B., & Botvinick, M. M. (2010). Decision making and the avoidance of cognitive demand. *Journal of Experimental Psychology: General, 139*(4), 665-682.

LaMantia, J., Rennie, W., Risucci, D., Cydulka, R., Spillane, L., Graff, L., et al. (1999). Interobserver Variability among Faculty in Evaluations of Residents Clinical Skills. *Academic Emergency Medicine, 6*(1), 38-44.

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87*(1), 72-107.

Lavie, N. (1995). Perceptual load as a necessary condition for selective attention. *Journal of Experimental Psychology: Human Perception and Performance, 21*(3), 451-468.

Levitin, D. J. (2002). *Foundations of cognitive psychology: core readings*: The MIT Press.

Lord, R. G., & Maher, K. J. (1990). Alternative information-processing models and their implications for theory, research, and practice. *The Academy of Management Review, 15*(1), 9-28.

Lurie, S., Mooney, C., & Lyness, J. (2009). Measurement of the general competencies of the Accreditation Council for Graduate Medical Education: a systematic review. *Academic Medicine, 84*(3), 301-309.

Mack, A. (2003). Inattentional blindness. *Current Directions in Psychological Science, 12*(5), 180-184.

Mack, A., Rock, I. (1998). *Inattentional Blindness.* Cambridge, MA: MIT Press.

Margolis, M., Clauser, B., Cuddy, M., Ciccone, A., Mee, J., Harik, P., et al. (2006). Use of the mini-clinical evaluation exercise to rate examinee performance on a multiple-station clinical skills examination: a validity study. *Academic Medicine, 81*(10), S56-S60.

Marois, R., & Ivanoff, J. (2005). Capacity limits of information processing in the brain. *Trends in Cognitive Sciences, 9*(6), 296-305.

Mayer, R.E. (2010). Applying the science of learning to medical education. *Medical Education, 44*(6), 543-549.

Melchers, K. G., Kleinmann, M., & Prinz, M. A. (2010). Do Assessors Have Too Much on their Plates? The Effects of Simultaneously Rating Multiple Assessment Center Candidates on Rating Quality. *International Journal of Selection and Assessment, 18*(3), 329-341.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review, 63*(2), 81-97.

Morgeson, F., & Campion, M. (1997). Social and cognitive sources of potential inaccuracy in job analysis. *Journal of Applied Psychology, 82*(5), 627-654.

Noel, G., Herbers, J., Caplow, M., Cooper, G., Pangaro, L., & Harvey, J. (1992). How well do internal medicine faculty members evaluate the clinical skills of residents? *Annals of Internal Medicine, 117*(9), 757-765.

Norcini, J., Blank, L., Duffy, F., & Fortna, G. (2003). The mini-CEX: a method for assessing clinical skills. *Annals of Internal Medicine, 138*(6), 476-481.

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

Pashler, H. Ed. (1998). *Attention*. East Sussex, UK: Psychology Press, Ltd.

Payne, J. W. (1980). Information processing theory: Some concepts and methods applied to decision research. *Cognitive processes in choice and decision behaviour,* (pp. 95-115). Watson, TS. USA. Lawrence Erlaum. Associates.

Pérez Moreno, E., Conchillo, Á., & Recarte, M. A. (2011). The Role of Mental Load in Inattentional Blindness. *Psicológica: Revista de metodología y psicología experimental, 32*(2), 255-278.

Regehr, G., Bogo, M., Regehr, C., & Power, R. (2007). Can we build a better mousetrap? Improving the measures of practice performance in the field practicum. *Journal of Social Work Education, 43*(2), 327-344.

Schulz, C. M., Skrzypczak, M., Schneider, E., Hapfelmeier, A., Martin, J., Kochs, E. F., et al. (2011). Assessment of subjective workload in an anaesthesia simulator environment: reliability and validity. *European Journal of Anaesthesiology, 28*(7), 502-505.

Shah, A. K., & Oppenheimer, D. M. (2008). Heuristics made easy: An effort-reduction framework. *Psychological Bulletin, 134*(2), 207-222.

Simons, D. (2000). Attentional capture and inattentional blindness. *Trends in Cognitive Sciences, 4*(4), 147-155.

Thompson, W. G., Lipkin, M., Jr., Gilbert, D. A., Guzzo, R. A., & Roberson, L. (1990). Evaluating evaluation: assessment of the American Board of Internal Medicine Resident Evaluation Form. *Journal of General Internal Medicine, 5*(3), 214-217.

Tsang, P. S., & Vidulich, M. A. (2006). Mental workload and situation awareness. In *Handbook of human factors and ergonomics. Third Ed. (ed. G. Salvendy)*. (pp. 243-268) Hoboken NJ: Wiley and Sons.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124-1131.

Van Merriënboer, J. J. G., & Sweller, J. (2010). Cognitive load theory in health professional education: design principles and strategies. *Medical Education, 44*(1), 85-93.

Wickens, C. D., & Carswell, C. (2012). Information processing. *Handbook of human factors and ergonomics* (pp. 111-149). Hobokenm NJ. USA. Wiley and Sons.

Williams, R., Klamen, D., & McGaghie, W. (2003). Cognitive, Social and Environmental Sources of Bias in Clinical Performance Ratings. *Teaching and Learning in Medicine, 15*(4), 270-292.

Young, G., Zavelina, L., & Hooper, V. (2008). Assessment of workload using NASA Task Load Index in perianesthesia nursing. *Journal of PeriAnesthesia Nursing, 23*(2), 102-110.

Yurko, Y. Y., Scerbo, M. W., Prabhu, A. S., Acker, C. E., & Stefanidis, D. (2010). Higher Mental Workload is Associated With Poorer Laparoscopic Performance as Measured by the NASA-TLX Tool. *Simulation in Healthcare, 5*(5), 267-271.

Zheng, B., Cassera, M. A., Martinec, D. V., Spaun, G. O., & Swanström, L. L. (2010). Measuring mental workload during the performance of advanced laparoscopic tasks. *Surgical Endoscopy, 24*(1), 45-50.

1.

Chapter 3 – "Global Rating Scale for the Assessment of Paramedic Clinical Competence"

Tavares, W., Boet, S., Theriault, R., Mallette, T., & Eva, K.W. (2012b). Global Rating Scale for the Assessment of Paramedic Clinical Competence. *Prehospital Emergency Care, 17*(1), 57-67.

Abstract

Objective: The aim of this study was to develop and critically appraise a global rating scale (GRS) for the assessment of individual paramedic clinical competence at the entry to practice level.

Methods: The development phase of this study involved task analysis by experts, contributions from a focus group and a modified Delphi process using a national expert panel to establish evidence of content validity. The critical appraisal phase had 2 raters apply the GRS, developed in Phase 1, to a series of sample performances from 3 groups: novice paramedic students (G1), paramedic students at the entry to practice level (G2), and experienced paramedics (G3). Using data from this process, we examined the tool's reliability within each group and tested the discriminative validity hypothesis that higher scores would be associated with higher levels of training and experience.

Results: The development phase resulted in a 7-dimension, 7-point adjectival GRS. The two independent blinded raters scored 81 recorded sample performances (n=25 in G1, n=33 in G2, n=23 in G3) using the GRS. For G1, G2 and G3, respectively, inter-rater reliability reached 0.75, 0.88 and 0.94. Intra-rater reliability reached 0.94 and the internal consistency ranged from 0.53 to 0.89. Rater differences contributed 0% - 5.7% of the total variance. GRS scores assigned to each group increased with level of experience, both using the overall rating (Means=2.3, 4.1, 5.0; $p<0.001$) and considering each dimension separately. Applying a modified borderline group method, 54.9% of G1, 13.4% of G2 and 2.9% of G3 were below the cut score.

Conclusion: The results of this study provide evidence that the scores generated using this scale can be valid for the purpose of making decisions regarding paramedic clinical competence.

Introduction

Paramedics provide emergency and non-emergency care to patients suffering diverse complex medical conditions and traumatic injuries. The level of clinical competence expected of paramedics has grown and like other health professions, lack of competence can adversely affect patient safety and outcomes. Educational institutions, employers, licensing bodies and/or regulators have a responsibility to ensure paramedic candidates entering the profession are ready for independent practice. Performance-based exams are an integral part of ensuring clinical competence [1]. Administering them well requires evidence of adequate reliability and validity [2].

In the field of paramedicine, assessment tools have generally taken the form of task specific checklists [3-5]. For example, the National Registry of Emergency Medical Technicians includes as part of its examination a series of task specific (often chronological) binary checklists describing performance expectations along with critical errors for each task (http://www.nremt.org). While checklists may be appropriate in some contexts (e.g. isolated procedural tasks), paramedic clinical competence, which includes both technical and non-technical elements as well as variations in process, may be challenging to identify and measure using checklists [6, 7]. Global rating scales (GRS) have rarely been considered for the assessment of paramedic competence [8, 9] despite their apparent necessity and associated advantages [7, 10, 11]. Global rating scales are subjective, but they have been demonstrated to better differentiate levels of experience when compared to checklists [7, 12, 13] although checklists may better differentiate between individuals within a novice group of examinees [14]. Recently, Martin et al., revealed considerable variability in the rate at which raters reported errors when observing videotaped performances of paramedic practice [9]. In general, however, there is a paucity of research evaluating the reliability and validity of assessment tools in paramedicine, and therefore, no gold standard (checklist or GRS) exists.

Kane (2006) describes validation as a process of evaluating proposed interpretations of data based on the scores generated from an instrument. This involves clearly stating intended interpretations, identifying assumptions and critically evaluating the assumptions associated with the measurement tool [15]. The lack of this type of research in paramedic settings raises concerns regarding the accuracy and defensibility of the performance-based assessments used. This study aimed to develop and critically appraise a generic global rating scale to measure individual paramedic clinical competence for summative 'entry to practice' decisions.

Methods

Ethical approval was provided by Centennial College in Toronto, Ontario, Canada (REB # 087) and informed consent was obtained from each participant. We structured our study in two phases: a development phase in which we ensured the construct of paramedic clinical competence was adequately represented, and an appraisal phase in which we critically evaluated intended interpretations using the GRS.

Study Design for Development Phase

The development phase of this study focused on ensuring that the rating scale adequately represented the target construct of primary care paramedic clinical competence. Local experts, a broad focus group, and then a national panel of experts representing a variety of stakeholders in

paramedic education, certification and employment engaged in an iterative process until there was evidence of agreement that the construct was adequately represented by the scale.

First, local experts (WT and two other faculty from Centennial College) engaged in task analysis aimed at identifying relevant behaviours in paramedic clinical practice through observation of various clinical cases completed both in simulation and actual clinical practice. This group then clustered the behaviours observed and ultimately identified specific dimensions so that a working prototype GRS could be prepared. Next, a focus group representing end users (i.e., educators, certifying bodies) and raters was assembled to evaluate the characteristics, items, definitions and language used in the working prototype. This focus group was given an opportunity to apply the working prototype to a series of sample video performances to identify concerns, issues, or gaps. Finally, a national expert panel engaged in a modified Delphi process to complete the development phase [16, 17]. This involved presenting experts with a sample list of dimensions with bulleted statements intended to define each (developed and refined in task analysis and focus group session), using a web-based survey tool. Experts were asked to rate the relevance of each element (i.e., dimension and descriptor, 7-point labels and descriptors) as it relates to the intended construct using a 4-point scale from 1=not relevant to 4=extremely relevant, rate their level agreement with regards to adequate representation of the construct using a 4-point scale from 1=strongly disagree to 4=strongly agree, and provide comments. Results were then shared with each expert panel independently in subsequent rounds.

Analysis

Local experts continued the task analysis until saturation (i.e., until no new distinct and relevant behaviours were identified). Clustering continued iteratively until all behaviours were organized, dimensions could be identified, and a working prototype GRS was prepared. The focus group engaged in open discussion facilitated by the principal investigator (WT) until saturation (i.e., no new changes / revisions were suggested). Finally, using the working prototype GRS, the modified Delphi process continued until consensus was reached (defined as 80% agreement) among national experts on all items, definitions, rating labels and rating label definitions.

Study Design for Critical Appraisal Phase

The critical appraisal phase of this study tested the hypotheses that (i) the dimensions listed in the prototype are distinct and adequately represent the construct of interest, (ii) individual paramedics can be consistently differentiated by raters using the GRS (i.e., that the tool is reliable), and (iii) higher scores would be empirically associated with higher levels of experience when using the GRS to rate paramedic clinical performance. To test these hypotheses we subjected the scale to a quasi-experimental design to evaluate internal structure, reliability and relationship to other variables. This involved first recording clinical performances by three distinct groups; *novice* (in training) paramedic students (Group 1), *entry to practice* (about to graduate) paramedic students (Group 2), and *experienced* paramedics (Group 3) who all completed the same case in simulation. The videos were coded, randomized and distributed to raters to score using the prototype GRS.

Participants and Scenario

Purposive sampling was used to recruit paramedic students for groups 1 (novice) and 2 (entry to practice) from a local paramedic program and to recruit participants for group 3 (experienced paramedics) from 5 different Emergency Medical Services in Southern Ontario, Canada. Our selection of these groups was based on evidence that suggests expertise develops as a result of a greater knowledge base [18], greater experience through supervised and unsupervised exposure to a variety of patients [18, 19] and more opportunity for deliberate practice [19]. This provided a range of competence that could then be used to test the scale's ability to differentiate between levels of performance.

Participants in each group were required to complete the same case in a high fidelity simulator. SimMan® **(**Laerdal Medical, Stavenger, Norway) was placed in a mock ambulance equipped with audio and video recording equipment. The simulated scenario involved an unstable cardiac patient with decreased level of consciousness who, at a predetermined marker, deteriorated to cardiac arrest over 9 minutes. Confederate "transfer company ambulance staff" were present as part of the simulation to provide a history to the examinee by answering his/her queries. Because the scenario was set at the "side of the roadway", as described in appendix A, participants also had to demonstrate awareness of surveying the scene and assuring the safety of those involved. The scenario used was similar to those included in traditional entry to practice assessment processes and required a broad range of technical and non-technical skills. It was based on an actual clinical case, piloted and refined using paramedic education and simulation experts, students and active paramedics. Provincial and national scope of practice guidelines informed the case development process [20-23]. Participants were instructed to assess and manage the clinical case to the best of their ability using any available equipment and resources. Content and performance expectations were carefully considered to ensure all groups (including Group 1) had sufficient knowledge and skill to complete the case to standard without identifying themselves as being at a particular level of training. We intentionally selected paramedics for group 3 with less than 5 years of experience to reduce heterogeneity and limit differences in appearance that could bias the ratings.

Sample Size Calculation

The primary outcome for this study was the ability of raters to differentiate between groups 1, 2 and 3 using the prototype global rating scale. With an estimated effect size of 0.8, which is widely accepted in education and psychology to indicate a large effect [24] and has been used in similar scale validation studies [25], a two-tailed α of 0.05 and β of 0.20, 25 participants were required per group for a total of 75 participants. Rounding up to account for potential attrition, we sought to enroll 85 participants. This sample size facilitated secondary outcome (e.g., item analysis, inter-rater reliability) analyses as well [26].

Rating Procedure

Videos were given a study code, randomly ordered to minimize potential confounders, and then distributed to 2-blinded independent raters (described below) who were asked to score the videos as they normally would for entry to practice decisions. Raters were allowed to take notes while observing the performance, but rewinding or pausing the video was not allowed in order replicate natural conditions as closely as possible. Raters were informed that the videos represented a collection of performances from a variety of clinicians, and that each was to be evaluated independently using the GRS. Prior to scoring, raters were provided with a brief

introduction to the rating scale and given instructions on how to apply the scale. One sample case from pilot recordings was used to allow raters to practice applying the scale before beginning data collection. Both raters and the principal investigator (WT) met following the initial practice rating to discuss any rating issues. The introduction, instructions, practice sessions and discussion took approximately 60 minutes. We intentionally limited rater training to evaluate outcomes under the most natural conditions. No attempt was made to calibrate the raters and only scale application issues were discussed. Two months following the initial rating of all videos, each rater was randomly assigned a subset of the videos to enable an evaluation of intra-rater reliability.

Analysis

The scores generated are expected to be used to make inferences regarding paramedic clinical competence. To test assumptions related to the scale's content, internal structure was analyzed via item analysis (i.e., internal consistency, inter-item and item total correlations) [27]. Variance attributable to raters, items, and the relevant interactions between those facets was determined and used to calculate reliability. Inter and intra-rater reliability was calculated using generalizability theory [28]. A modified borderline group method [29-31] for establishing cut scores was also applied and reported using descriptive statistics. This involved having raters judge candidate performance using a 7-point adjectival scale that was included at the end of the GRS. They were asked to rate each candidate's overall performance as either unsatisfactory (1=unsafe, 2=unsatisfactory, 3=poor/weak) or satisfactory (4=marginal, 5=competent, 6=highly competent, 7=exceptional). Prior to the rating task, raters were informed scores of 3 or 4 on this overall category would be considered the borderline group. Scores assigned on the 7 construct specific dimensions for that cohort of candidates were aggregated to establish a cut score by dimension. Finally, using scores from each of the three groups, we tested the assumption that higher scores are related to higher levels of experience using Analysis of Variance (ANOVA). All data were analyzed using SPSS Version 19 and Generalizability software (G-string Version IV). The level of significance was set at p=.05 (two-tailed).

Results

Development phase

Task Analysis and Item Development

The development phase involved having experts from a paramedic program conduct a task analysis [32] using multiple simulation-based paramedic clinical performances and actual clinical cases. Experts identified 257 observable behaviours from a variety of contexts and then iteratively arranged the behaviours into clusters relevant to paramedic practice. Additional performance observations continued to determine sufficiency of the list and/or the need for further refinements to the clusters. Eight dimensions in total were identified: situation awareness, history gathering, patient assessment, decision-making, implementation, resource utilization, communication and procedural skills. Using the behaviours identified during the task analysis, descriptors with examples for each were attached to each dimension. An initial working prototype GRS including the 8 dimensions and 7-point adjectival scales was created. A 7-point scale was selected to facilitate reliability without creating levels raters would have difficulty differentiating between [26, 33]. Rating labels, with definitions for each (based on

practice standards, patient safety, and readiness for independent practice or progression) anchored each of the 7 points.

Focus Group

Next, a focus group of 17 practicing paramedic clinicians who were also practicing educators and assessors from 5 different emergency medical services in southern Ontario, Canada contributed to the refinement of the scale. Raters reviewed and approved the list of dimensions assembled, the definitions associated with each dimension, the rating labels selected and their definitions. After having an opportunity to apply the scale to 2 pre-recorded videos of paramedic simulations, the dimension "implementation" was identified as a source of disagreement regarding its distinction from other dimensions. The focus group along with the researchers elected to retain the dimension for the national expert panel.

National Expert Panel

Nine experts from 5 provinces across Canada participated in a modified Delphi process [16, 17]. Experts were selected based on their individual experience in paramedicine and unique perspective relative to the rating scale's intended application (i.e., entry to practice decisions). For example, some were responsible for graduating paramedic candidates (n=5), while others were responsible for employment (n=2), and still others for certification (n=2). All were practicing experienced educators (n=9, median of 10 years in paramedic education), researchers (n=3), or active paramedics (n=8, median of 15 years in clinical practice). Round 1 of the Delphi achieved consensus (>80% agreement) on all dimensions, rating labels and definitions except for the dimension "implementation". Following round 1, bulleted statements were converted to general descriptions for each dimension and suggestions for revisions were implemented or shared with the group for consensus prior to round 2. In Round 2 a revised GRS was distributed and achieved consensus on all levels with the exception of the "implementation" dimension. Similar to the focus group session, the expert panel disagreed regarding its distinction from other dimensions and its inherent inclusion in each. Based on the feedback from the focus group, and results of the Delphi process, the dimension "implementation" was eliminated from the rating scale and the Delphi process discontinued. A copy of the final GRS is included in appendix B.

Critical appraisal

Group Participants and Raters

Participants for each group were enrolled between January and May of 2011. Eighty-five participants were enrolled. Twenty-five novice paramedic students (17 males, 8 females) in Group 1, 36 entry to practice students (19 males, 17 females) in Group 2 and 24 active paramedics (14 male, 10 females) in Group 3. Three videos in group 2 were discarded due to technical difficulties. One video in group 3 was discarded once a participant disclosed he was not an active paramedic. Of the remaining 23 paramedics in Group 3, the mean years of experience was 2.4 and 6 different paramedic services and 6 different paramedics programs in Ontario Canada were represented. A total of 81 videos, each lasting 9 minutes, were submitted to 2 raters for scoring. The initial rating procedure was completed over a 1 month period. The subsequent rating procedure (used to calculate intra-rater reliability) involved rating a random

selection of 30 of the 81 videos. This second rating took place 2 months following the initial rating task and was completed over a 2 week period.

Two raters were selected from 2 different paramedic programs in Ontario, Canada. Together the raters averaged 11 years experience as paramedic educators, 22 years as paramedics and 13 years evaluating clinical performances. All videos were scored on 7 dimensions, and given an "overall" performance rating by each rater.

Reliability

Reliability analyses were conducted on each group independently to avoid artificially inflating the heterogeneity in the videos. The proportion of variance attributable to rater differences in Groups 1, 2, and 3, respectively, was 0.07%, 0%, and 5.72% of group total variance. All variance components are illustrated in Table 1.

Using participant as the facet of differentiation and items as the facet of generalization, internal consistency was calculated and found to be .89, .71 and .53 for Groups 1, 2, and 3, respectively. Inter-item correlations ranged from .62 to .93 and item-total correlations ranged from .74 to .92. Individual inter-item and item-total correlations are provided in Table 2 along with the correlation between each item and the overall rating assigned.

The inter-rater reliability for Groups 1, 2 and 3 reached .75, .88 and .94, respectively. Intra-rater reliability was calculated using the scores assigned to the 30 randomly selected videos and reached .94. The reliability for each dimension considered independently, ranged from .54 (communication) to .83 (decision making) within Group 2 (i.e., those selected from the target population). Individual g-coefficients for each dimension are reported in Table 3.

Relationship to Other Variables

To test for evidence of discriminative validity, using all dimensions, a one-way ANOVA was performed using group as the independent variable and average score as the dependent variable. The effect of group was found to be statistically significant both based on overall scores ($F(2,78) = 29.5$, $p = <.001$) and for each individual dimension (see Table 4). Making pairwise comparisons, the differences between the means across group aligned with expectations in 23 out of 24 instances, consistency that can be expected to occur less than 0.1% of the time according to binomial probability theorem (i.e., $p<0.001$). The one reversal (Group 2 > Group 3 in the Communication skills dimension) was slight with an observed difference of 0.06.

We applied the modified borderline group method to each dimension to evaluate the relationship between failure rate and group assignment and found the highest failure rates in Group 1, followed by Group 2, and then Group 3. Results are provided in Table 5.

Discussion

Kane (2006) describes validation as a process of generating an *interpretive argument* in which proposed interpretations are clearly stated (such as higher scores on the GRS are indicative of a higher level of experience) and then critically evaluated for plausibility and coherence [15]. There are numerous assumptions between the observation of performance and the final decision regarding competence that must be identified and evaluated if the interpretation based on scores generated is to be considered defensible. Using this framework, we proposed that the scale

would be used to make inferences regarding paramedic clinical competence at the entry to practice level. We then identified construct representation as the first assumption to be tested and hypothesized that higher scores would align with higher levels of experience. The results of a development and critical appraisal process included in this study, suggest that this GRS can be implemented in a way that provides reasonable reliability and capacity to differentiate both between groups and between individuals within group and, therefore, enables inferences regarding paramedic clinical competence.

The development phase and early stages of the critical appraisal phase of this study were aimed at evaluating the adequacy of construct representation. This content validation process involved clinicians, educators and experts in the field of paramedicine collectively and iteratively ensuring an appropriate focus of the GRS. This involved detailed task analyses (i.e., observation of clinical performance using a wide variety of cases) in simulation and clinical settings, a large focus group of clinicians who were also educators and raters, and a national expert panel representing a variety of stakeholders responsible for making decisions regarding independent paramedic practice. All were implemented using rigorous item construction rules and processes to devise and refine items. This resulted in a 7 dimension GRS: situation awareness, history gathering, patient assessment, decision-making, resource utilization, communication and procedural skill.

Once the development phase was complete and a GRS created, we subjected the scale to a quasi-experimental design. We recruited a range of clinicians (i.e., paramedic students at different levels of training and experienced paramedics) to complete the same case in a simulation setting, then had raters, blinded to group, observe and rate the clinical performances. Data collected from these processes allowed us to conduct item and reliability analyses and to evaluate the relationship of the scores to the training/experience level of participants, a practice that has been successfully applied in other similar studies [6, 25, 34, 35].

The high inter-item correlations observed suggest that the items, despite representing diverse dimensions, were possibly measuring a single construct. These high inter-dimension correlations reinforce Lurie, Mooney and Lyness' findings, which suggest that raters have difficulty differentiating between dimensions [37]. This may be an indication that, psychologically, raters form Gestalt categorical judgments about candidates/trainees as part of impression formation (i.e., a halo effect) [38] perhaps due to difficulty tracking multiple dimensions simultaneously [33]. Still, the scale demonstrated evidence of inter-rater and intra-rater reliability, with minimal error attributed to rater. This may be in part due to deliberate efforts to align the label definitions (e.g., ready for independent practice) with the manner in which clinical supervisors naturally conceive of trainees' progress, a strategy that Crossley, Johnson, Booth and Wade has shown can improve rating practice [39]. Finally, the scores generated were significantly different between groups. These results strengthen confidence in the inferences made based on scores generated using this scale. That is, they strengthen the interpretive argument and suggest this GRS can be used for the assessment of paramedic clinical competence during entry to practice assessment processes.

These findings add to the broader health professional literature suggesting the use of global rating scales to be a suitable measurement strategy. Crossley and Jolly claim that assessors judge performance more consistently and discriminatingly when not tied to process level observation

(i.e., reducing complex clinical performance to a series of individual steps)[40]. This may help explain why the global ratings used in our study appear to be more reliable than those used by Han et al. [14] despite the fact that we examined reliability within a group of novice practitioners just as they did. Our goal was to evaluate paramedic clinical competence at the entry to practice level. The variations in appropriate performance that can exist among clinicians at this level may not be amenable to process level assessment (e.g., checklists). For instance, in making judgments regarding clinical performance, checklists may facilitate the assessment of occurrence (i.e., whether or not particular behaviours were present) but GRS may be more suitable for considering quality (i.e., how good the performance is) and suitability (i.e., whether or not the performance was good enough for entry to practice) [12]. Rather, outcome or structure level assessments and definitions [40], which are included in this GRS may be better suited.

Limitations

As always, there are limitations associated with this study. First, we used only one unscripted case (a medical cardiac patient). This tells us that performance can be reliably differentiated, but it limits external validity and prevents us from determining the extent to which individuals' general ability is captured by a single application of the scale. In terms of external validity, whether similar results would be found when assessing candidates attending to a trauma victim, for example, requires further study. That said, the case involved a number of interactions (e.g., communicating with unhelpful staff, integrating available resources, a need for selecting appropriate assessment strategies) that would likely be applicable in a variety of patient encounters. Further, the development phase of this study, included as part of each phase (i.e., task analysis, focus group and expert panel) the consideration and inclusion of a variety of contexts. Still, future studies will need to apply the GRS to other contexts (e.g., using different cases, in actual clinical competence exams) to assess the generalizability of the results reported here. With respect to drawing inferences about individual paramedics' general level of competence, the universality of context specificity [40] requires that research be done to determine how many times the GRS needs to be applied (i.e., how many cases need to be observed) to generate stable representation of an individual's competence level. Second, we had 2 independent expert raters score all 81 videos. The level of expertise of the raters as well as the repetition may have contributed to the results (e.g. high inter-rater reliability, low error variance) though it is worth noting that raters did not compare notes over the course of completing their assignment and, hence, were as prone to drifting apart in their perceptions as they were to come to a mutual understanding of how well individual participants performed. Finally, in critically appraising the GRS, we selected 3 groups (year 1 paramedic students, year 2 entry to practice students, and experienced paramedics) to test the scales' discriminative validity. The heterogeneity of these groups could be argued to have had an effect on our study results, which demonstrate an ability to differentiate between groups. The scale was designed to support decisions for entry to practice. The groups selected provide a range, therefore, around the population of interest. Further, within each group we were able to identify a range of performance levels (based on the high level of participant variance within group) and reasonable reliabilities were observed in all groups including group 2, our intended target. What cut scores should be used in actual practice will depend on many factors including the way in which the data are to be used (e.g., for formative or summative purposes and the stakes involved in the decision to be made). Interested readers are directed to other sources for a more comprehensive

treatment of assessment strategies and standard setting in health professions education [2, 12, 41-43].

Conclusions

Paramedic program educators, employers, certifying and/or licensing bodies all have a responsibility to ensure those who are ultimately given access to independent paramedic practice are indeed competent. This requires the use of appropriate process and measurement tools with sufficient scientific evidence to support inferences or interpretations based on the scores generated. This study provides support for use of our rigorously developed GRS in practice by demonstrating evidence of content validity, sound psychometric properties, limited construct irrelevant variance and an ability to differentiate between levels of performance. Applied in the proper context, this scale could help strengthen decisions regarding paramedic clinical competence. Additional research is recommended to further support this interpretive argument, especially in other contexts.

Declaration of Interest

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

Acknowledgements

The authors would like to thank all the students, paramedics and experts who agreed to participate in this study. The authors would also like to thank (from east to west) the Emergency Health Service of Nova Scotia; Centennial College Science and Technology Centre / University of Toronto, ORNGE, Sunnybrook Centre for Prehospital Medicine, Georgian College, Lambton College of Applied Arts and Technology, and York Region Emergency Medical Services of Ontario; Red River College of Manitoba; Saskatchewan Institute of Applied Science and Technology; Lakeland College, University of Alberta and Professional Medical Associates of Alberta; Justice Institute of British Columbia, and the Society for Prehospital Educators of Canada (SPEC) for their contributions.

Financial Support

This study was funded by the Applied Research and Innovation Centre at Centennial College in Toronto, ON, Canada.

Table 1 – Variance components and percentage of total variance by group.

Effect	G1 VC	% of Total Var.	G2 VC	% of Total Var.	G3 VC	% of Total Var.
person	0.89	41.1%	1.05	47.8%	0.76	32.2%
rater	0.00	0.1%	0.00	0.0%	0.14	5.7%
item	0.07	3.2%	0.00	0.0%	0.06	2.4%
person x rater	0.11	5.1%	0.24	11.1%	0.34	14.5%
person x item	0.25	11.5%	0.17	7.9%	0.10	4.2%
rater x item	0.24	11.3%	0.23	10.3%	0.19	8.1%
person x rater x item	0.59	27.1%	0.50	22.9%	0.77	32.78%
Total Variance	2.16	100%	2.19	100%	2.35	100%

G1 = Group 1 novice level paramedic students, G2 = Group 2 entry to practice level paramedic students, G3 – Group 3 experienced active paramedics; VC = variance components; Var. = variance.

Table 2: Inter-item and item total correlations using data from all 3 groups.

Dimension		SA	HG	PA	DM	RU	COM	Item-Total Correlation	Correlation with "Overall" rating
Situation Awareness	(SA)							.93	.95
History Gathering	(HG)	.71						.74	.74
Patient Assessment	(PA)	.93	.69					.89	.91
Decision Making	(DM)	.92	.70	.88				.92	.95
Resource Utilization	(RU)	.81	.67	.77	.79			.85	.84
Communication	(COM)	.69	.63	.62	.66	.76		.74	.75
Procedural Skill	(PS)	.85	.66	.81	.89	.75	.67	.88	.92

Table 3: Inter-rater reliability for each dimension calculated using generalizability theory and Group 2 data.

Dimension	G-Coefficient
Situation Awareness	.83
History Gathering	.64
Patient Assessment	.81
Decision Making	.84
Resource Utilization	.60
Communication	.54
Procedural Skill	.67

G-coefficient = generalizability coefficient.

Table 4: Descriptive statistics and ANOVA results by dimension.

	Descriptives					**Analysis of Variance**			
Dimension	Group	Mean	Std. Dev.	95% CI		df	Mean Square	F	P Value
Situation Awareness	1	2.36	1.24	1.85	2.87	2,78	52.49	32.44	.00
	2	4.21	1.39	3.72	4.70				
	3	5.26	1.13	4.77	5.75				
History Gathering	1	3.10	1.28	2.57	3.63	2,78	10.94	9.03	.00
	2	4.03	1.07	3.65	4.41				
	3	4.39	0.90	4.00	4.78				
Patient Assessment	1	2.16	1.02	1.74	2.58	2,78	36.00	25.09	.00
	2	3.42	1.32	2.96	3.89				
	3	4.61	1.20	4.09	5.13				
Decision Making	1	2.28	1.30	1.74	2.82	2,78	57.11	30.00	.00
	2	4.14	1.43	3.63	4.64				
	3	5.33	1.39	4.72	5.93				
Resource Utilization	1	2.78	1.00	2.37	3.19	2,78	26.06	22.34	.00
	2	3.98	1.14	3.58	4.39				
	3	4.85	1.08	4.38	5.32				
Comm.	1	3.42	1.19	2.93	3.91	2,78	8.06	5.14	.01
	2	4.41	1.11	4.01	4.80				
	3	4.35	1.49	3.70	4.99				
Procedural Skill	1	2.78	1.44	2.18	3.38	2,78	35.46	21.40	.00
	2	4.17	1.33	3.70	4.64				
	3	5.20	1.02	4.75	5.64				
Overall	1	2.30	1.22	1.80	2.80	2,78	46.27	29.48	.00
	2	4.05	1.38	3.56	4.54				
	3	5.02	1.08	4.55	5.49				

Std. Dev. = standard deviation; CI = confidence interval; df = degrees of freedom

Table 5: Percentage of individuals below cut score as defined by the Modified Borderline Group Method (by group and by dimension).

Dimension	SA	HG	PA	DM	RU	COM	PS	Mean
Group 1	56%	56%	56%	64%	52%	48%	52%	54.9%
Group 2	12%	18%	21%	1%	12%	15%	15%	13.4%
Group 3	0%	1%	0.3%	0.3%	0.3%	18%	0.3%	2.9%

G1 = Group 1: novice level paramedic students, G2 = Group 2: entry to practice level paramedic students, G3 – Group 3: experienced active paramedics

Appendix A – Case summary.

OVERVIEW

This case involved a paramedic (i.e., the candidate) working alone and responding to the side of a roadway for a patient with a decreased level of consciousness who was in the rear of a transfer company ambulance[1]. According to transfer company staff, the patient's condition began with severe shortness of breath secondary to congestive heart failure that progressed into lethal arrhythmia and eventually cardiac arrest. The transfer company staff are "on scene" (i.e., in the rear of the transfer company vehicle with the patient) arguing over who is responsible for the current predicament.

CALL INFORMATION

Call for a 75 year old male/female with shortness of breath.

CASE DETAILS

The patient [mannequin] presented initially as responding only to painful stimuli with moans, was diaphoretic and tachypneic. His presenting rhythm was ventricular tachycardia. Presenting vital Signs were: BP: 68/48, HR: 190 (ventricular tachycardia), RR: 30 s/r, (crackles throughout). BS = 8.8. Patient had a history of Alzheimer's, coronary artery disease, two previous MIs, CHF, CVA [no lasting deficits], hypertension, diabetes, and high cholesterol. The medication list included Aricept, Metoprolol, Digoxin, Lisinopril, Glucophage, Atorvastatin, and was allergic to morphine.

The transfer company crew members remain "on scene" for the duration of the interaction and are available to provide history verbally (no patient chart) in response to the candidate's requests. The candidate must use effective communication skills to elicit the information as the two transfer company staff continues to argue, be disruptive and are reluctant to assist. However, one of the two transfer company staff is able to serve as a resource to the candidate.

At the 5-minute point in the case, the patient becomes Vital Signs Absent.

[1] In Canada, some unregulated private companies may provide transfer services to patients. Generally, these unregulated transfer companies may not be held to the same standard as fully regulated ambulance services, including staff qualifications. Further, unregulated transfer companies, are not authorized to transport patients directly to emergency departments.

Appendix B – Global Rating Scale for the Assessment of Paramedic Clinical Competence

PARAMEDIC GLOBAL RATING SCALE

Candidate: _____ Rater: _____
Date: _____

Case Description:
_____

Rating Label **Definition**

1=Unsafe ……Not performed as required. Performance compromised patient care / safety; serious remediation is required, unsuitable for supervised practice or progression.

2=Unsatisfactory ….. Performance indicated cause for concern. A potential for compromised patient care / safety exists; considerable improvement is needed. Not ready for supervised practice or progression.

3=Poor / Weak …… Inconsistently performed, and/or performance does not meet the standard, improvement is needed. More training / practice is needed before consideration for supervised practice or progression.

4=Marginal ……..…. Occasionally performance is to standard, and/or performance meets minimum standards, improvement is recommended; suitable for supervised practice or progression with some remediation.

5=Competent …….. Often performed to standard, and/or performance is safe and to standard. Some areas could be improved. Ready for independent practice or progression with only minor concerns if any.

6=Highly Competent…Consistently performs to standard, and/or performance is safe and to standard. Occasionally exceeds the standard. Little improvement needed if any; ready for independent practice or progression.

7=Exceptional …… Consistently demonstrates a high standard of performance, and/or consistently exceeds the standard enhancing patient safety; could be used as a positive example for others; highly recommended for independent practice or progression.

Situation Awareness	1	2	3	4	5	6	7
	UNSAFE	**UNSAT**	**POOR/WEAK**	**MARGINAL**	**COMPETENT**	**HIGHLY COMPETENT**	**EXCEPTIONAL**

Refers to the individual's overall ability to consider and integrate environmental, scene, resources and patient condition cues into the overall interaction, management and safety plan. This includes observing whole environment (all available data sources), anticipating likely events, discriminating between relevant and irrelevant data and avoiding tunnel vision (inappropriately focusing on elements to the exclusion of others). The individual is expected to

demonstrate examples of situation awareness throughout the interaction and updating actions as necessary.

History Gathering	1	2	3	4	5	6	7
	UNSAFE	**UNSAT**	**POOR/WEAK**	**MARGINAL**	**COMPETENT**	**HIGHLY COMPETENT**	**EXCEPTIONAL**

Refers to the individual's overall ability to effectively and thoroughly gather an appropriate history (includes history of present illness and medical history) which is organized, appropriately structured, timed and focused according the clinical situation and level of urgency (context). This includes interpreting and evaluating findings while discriminating between relevant and irrelevant findings. Also, refers to a demonstrated ability to include a consideration for differential diagnosis, while working toward a working diagnosis.

Patient Assessment	1	2	3	4	5	6	7
	UNSAFE	**UNSAT**	**POOR/WEAK**	**MARGINAL**	**COMPETENT**	**HIGHLY COMPETENT**	**EXCEPTIONAL**

Refers to the individual's overall ability to select and perform a physical exam and investigation of signs and/or symptoms that is organized and appropriate given the clinical situation and level of urgency. This includes interpreting and evaluating findings while discriminating between relevant and irrelevant findings. Also, refers to a demonstrated ability to continue appropriate reassessment / detailed assessment as needed. Finally, this also includes a consideration for differential diagnosis, while working toward a working diagnosis.

Decision Making	1	2	3	4	5	6	7
	UNSAFE	**UNSAT**	**POOR/WEAK**	**MARGINAL**	**COMPETENT**	**HIGHLY COMPETENT**	**EXCEPTIONAL**

Refers to the individuals overall ability to select an appropriate, safe, and effective management plan and/or strategy. Decisions should be based on and supported by findings, consideration of risks, benefits and differential diagnosis. This involves having adequate information for decisions made (i.e., avoiding premature closure) and ensuring decisions are appropriately prioritized, and timed. This also includes selecting an appropriate management device, method, or technique based on evidence (i.e., situation awareness, patient condition, resources etc) and context.

Resource Utilization	1	2	3	4	5	6	7
	UNSAFE	**UNSAT**	**POOR/WEAK**	**MARGINAL**	**COMPETENT**	**HIGHLY COMPETENT**	**EXCEPTIONAL**

Refers to the individual's overall ability to identify and use resources effectively to accomplish goals and maximize care. This includes the delegation of tasks, the coordination of efforts, selecting appropriate members (e.g., allied agencies, patients etc) for a given task, ensuring

effectiveness and requesting additional resources as necessary. This also includes ability to function as a team with appropriate leadership.

Communication	1	2	3	4	5	6	7
	UNSAFE	**UNSAT**	**POOR/WEAK**	**MARGINAL**	**COMPETENT**	**HIGHLY COMPETENT**	**EXCEPTIONAL**

Refers to the individuals overall ability to clearly and accurately exchange information with the team, patient and/or bystander for optimal patient care and team effectiveness. This includes the use of concise and appropriate language, ensuring statements are directed at appropriate individuals and that messages are heard / received (i.e., closes the loop). This also includes demonstrating effective listening skills, demonstrating empathy, responding appropriately to statements by the team, patient or bystander. Actions are appropriately communicated with team, patient and bystander. Verbal and non-verbal are appropriate and congruent.

Procedural Skill	1	2	3	4	5	6	7
	UNSAFE	**UNSAT**	**POOR/WEAK**	**MARGINAL**	**COMPETENT**	**HIGHLY COMPETENT**	**EXCEPTIONAL**

Refers to the individuals overall ability to complete psychomotor or procedural skills or tasks effectively, appropriately and to standard. This involves a familiarity with equipment used, ensuring appropriate and safe application while completing tasks to standard and avoiding commission or omission errors. This also involves adaptability to failures / problems (as necessary) and ensuring team, patient and bystander safety while performing these procedures; includes appropriate execution, properly sequenced, and evaluating / reevaluating effectiveness.

Overall Clinical Performance						
UNSATISFACTORY			SATISFACTORY			
1	2	3	4	5	6	7

References:

1. Miller G. The assessment of clinical skills/competence/performance. Academic Medicine 1990;65(9):S63-S67.
2. Brennan RL. Educational measurement. 2006: Praeger Pub Text.
3. Regener H. A proposal for student assessment in paramedic education. Medical Teacher 2005;27(3):234-241.
4. Lammers, R.L., Byrwa, M.J., Fales, W.D., Hale, R.A. Simulation-based assessment of paramedic pediatric resuscitation skills. Prehospital Emergency Care, 2009. 13(3): p. 345-356.
5. Studnek, J.R., Fernandez, A.R., Shimber, B., Garifo, M., Correll, M. The Association Between Emergency Medical Services Field Performance Assessed by High-fidelity Simulation and the Cognitive Knowledge of Practicing Paramedics. Academic Emergency Medicine, 2011. 18(11): p. 1177-1185.
6. Hodges, B., Regehr, G., McNaughton, N., Toberius, R., Hanson, M. OSCE checklists do not capture increasing levels of expertise. Academic Medicine, 1999. 74(10): p. 1129-1134.
7. Regehr, G., MacRae, H., Reznick, R., Szalay, D. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. Academic Medicine, 1998. 73(9): p. 993-997.
8. von Wyl, T., Zuercher, M., Amsler, F., Walter, B., Ummenhofer, W. Technical and non-technical skills can be reliably assessed during paramedic simulation training. Acta Anaesthesiologica Scandinavica, 2009. 53(1): p. 121-127.
9. Martin, M., Hubble, M.W., Hollis, M., Richards, M.E. Interevaluator Reliability of a Mock Paramedic Practical Examination. Prehospital Emergency Care, 2012. Vol. 16, No. 2, p. 277-283.
10. Hodges, B., McIlroy, J.H. Analytic global OSCE ratings are sensitive to level of training. Medical education, 2003. 37(11): p. 1012-1016.
11. Martin, J., Regehr, G., Reznick, R., MacRae, H., Murnaghan, J., Hutchinson, C. Objective structured assessment of technical skill (OSATS) for surgical residents. British Journal of Surgery, 1997. 84(2): p. 273-278.
12. Swanwick, T., Understanding medical education. 2010: Wiley Online Library.
13. Govaerts, M.J.B., van der Vleuten, C.P.M., Schuwirth, L.W.T. Optimising the reproducibility of a performance-based assessment test in midwifery education. Advances in Health Sciences Education, 2002. 7(2): p. 133-145.
14. Han, J., Kreiter, C.D., Park, H., Ferguson, K.J. An experimental comparison of rater performance on an SP-based clinical skills exam. Teaching and learning in Medicine, 2006. 18(4): p. 304-309.
15. Kane, M. Validity. In *Educational Measurement*, Editor: Brennan R.L., 2006. 4. p. 17–64.

16. Morgan, P., Lam-McCulloch, J., Herold-McIlroy, J., Tarshis, J. Simulation performance checklist generation using the Delphi technique. Canadian Journal of Anesthesia, 2007. 54(12): p. 992-997.

17. de Villiers, M., de Villiers, P., Kent, A. The Delphi technique in health sciences education research. Medical Teacher, 2005. 27(7): p. 639-643.

18. Graber, M., Educational strategies to reduce diagnostic error: can you teach this stuff? Advances in Health Sciences Education, 2009. 14: p. 63-69.

19. Ericsson, K., Krampe, R., Tesch-Römer, C. The role of deliberate practice in the acquisition of expert performance. Psychological review, 1993. 100(3): p. 363-406.

20. Paramedic Association of Canada, National Occupational Competency Profile. 2001, Paramedic Association of Canada.

21. Ontario Ministry of Health, Emergency Health Services Branch. BLS Standards of Care. Vol. 2.0. 2007: Publications Ontario. Toronto, ON Canada

22. Ontario Ministry of Health, Emergency Health Services Branch. Advanced Life Support Standards. 2007, Publications Ontario. Toronto, ON Canada

23. Ontario Ministry of Training Colleges and Universities, Paramedic Program Standards. 2008, Queens Printer for Ontario: Toronto, ON Canada.

24. Cohen, J., Statistical Power Analysis for the Behavioural Sciences. 1977, New York, NY: Academic Press.

25. Kim, J., Neilpovitz, D., Cardinal, P., Chiu, M., Clinch, J. A pilot study using high-fidelity simulation to formally evaluate performance in the resuscitation of critically ill patients: The University of Ottawa Critical Care Medicine, High-Fidelity Simulation, and Crisis Resource Management I Study. Critical care medicine, 2006. 34(8): p. 2167-2174.

26. Streiner, D., Norman, G. Health measurement scales: a practical guide to their development and use. Fouth ed. 2008, New York: Oxford Univeristy Press.

27. Cook, D.A., Beckman, T.J. Current concepts in validity and reliability for psychometric instruments: theory and application. The American Journal of Medicine, 2006. 119(2): p. 166.e7-166.e16.

28. Brennan, R., Generalizability theory. Educational Measurement: Issues and Practice, 1992. 11(4): p. 27-34.

29. Wilkinson, T.J., Newble, D.I., Frampton, C.M. Standard setting in an objective structured clinical examination: use of global ratings of borderline performance to determine the passing score. Medical education, 2001. 35(11): p. 1043-1049.

30. Humphrey-Murto, S., MacFadyen, J.C. Standard setting: a comparison of case-author and modified borderline-group methods in a small-scale OSCE. Academic Medicine, 2002. 77(7): p. 729-732.

31. Cizek, G.J., Setting performance standards: Concepts, methods, and perspectives. 2001. Mahwah, NJ: Lawrence Erlbaum Associates Inc. Publishers.

32. Kirwan, B.E., Ainsworth, L.K. A guide to task analysis. 1992: Philedelphia, PA: Taylor & Francis.

33. Tavares, W. & Eva, K.W. Exploring the Impact of Mental Workload on Rater-Based Assessments. Advances in Health Sciences Education, 2012. 18(2): p. 291-303.

34. Holmboe, E.S., Huot, S., Chung, J., Norcini, J., Hawkins, R.E. Construct Validity of the MiniClinical Evaluation Exercise (MiniCEX). Academic Medicine, 2003. 78(8): p. 826-830.

35. Goff, B.A., Nielsen, P.E., Lentz, G.M., Chow, G.E., Chalmers, R.W., Fenner, D., Mandel, L.S. Surgical skills assessment: a blinded examination of obstetrics and gynecology residents. American journal of obstetrics and gynecology, 2002. 186(4): p. 613-617.

36. Winckel, C.P., Reznick, R., Cohen, R., Taylor, B. Reliability and construct validity of a structured technical skills assessment form. The American journal of surgery, 1994. 167(4): p. 423-427.

37. Lurie, S., Mooney, C., Lyness, J. Measurement of the general competencies of the Accreditation Council for Graduate Medical Education: a systematic review. Academic Medicine, 2009. 84(3): p. 301-309.

38. Gingerich, A., Regehr, G., Eva, K.W. Rater-Based Assessments as Social Judgments: Rethinking the Etiology of Rater Errors. Academic Medicine, 2011. 86(10): p. S1-S7.

39. Crossley, J., Johnson, G., Booth, J., Wade, W. Good questions, good answers: construct alignment improves the performance of workplace based assessment scales. Medical education, 2011. 45 (6): p. 560-569.

40. Crossley, J., Jolly, B. Making sense of work-based assessment: ask the right questions, in the right way, about the right things, of the right people. Medical education, 2012. 46(1): p. 28-37.

41. Cizek, G.J. Bunch, M.B. Standard setting: A guide to establishing and evaluating performance standards on tests. 2007: Thousand Oaks, CA. SAGE Publications.

42. Downing, S., Yudkowsky, R. Assessment in Health Professions Education. 2009: New York, NY. Taylor & Francis.

43. Eva, K., Assessment Strategies in Medical Education, in Medical Education: State of the Art, Salerno-Kennedy, R., O'Flynn, S. Editors. 2010, Nova Scotia Publishers, Inc.: Halifax.

Chapter 4 – The Impact of Rating Demands on Assessments of Clinical Competence

Abstract

Purpose:
The objective of this study was to evaluate the impact of increasing rating demands on rater based assessments of clinical competence.

Methods:
Participants were randomly assigned to one of four conditions (in a 2x2 factorial design) and asked to rate 3 pre-recorded unscripted clinical encounters illustrating 2 levels of performance (high, medium, low). One factor was the number of dimensions participants were asked to rate: 7 vs. 2. The other involved the requirement (or lack thereof) to conduct additional extraneous, but ecologically valid, tasks by attending to patient status and the activity of additional individuals observable on video. Outcome measures included number of dimension relevant behaviours identified, ability to discriminate between levels of performance, and inter-rater reliability.

Results:
Using the 2 dimensions common to both groups, ANOVA revealed a significant main effect of the number of dimensions included in the scale on the number of relevant behaviours identified: Participants in the 2D group identified more features than those in the 7D group. Both groups were able to differentiate between levels of performance, but post hoc analyses revealed significance on all pairwise comparisons in the 2D group and not in the 7D group. Inter-rater reliability increased from .45 in the 7D group to .70 when participants were required to consider only 2 dimensions.

Conclusions:
The results of this study provide preliminary evidence that requiring raters to consider a greater number of dimensions can decrease (a) the number of dimension relevant behaviours identified, the capacity to discriminate between levels of performance, and (c) inter-rater reliability.

Introduction

Accurate assessment of clinical competence continues to challenge health professional education (Huwendiek et al., 2010). One such challenge derives from the fact that clinical competence is complex, abstract and not directly measureable. As such, it must be inferred from behaviours demonstrated by candidates with the consequence that assessment of many aspects of clinical competence is largely dependent on rater judgment. However, rater-based assessments have been identified as fallible and problematic. Observed challenges include difficulty discriminating between dimensions of performance, persistence of common rater biases and poor inter-rater reliability (Downing, 2004, 2005; Haber and Avins, 1994; Lurie et al., 2009; Thompson et al., 1990; Williams et al., 2003; Yeates et al., 2012).

Until recently, common explanations for rater difficulties have mainly been grounded in issues related to insufficient rater training and/or scale development (Green and Holmboe, 2010; Lurie et al., 2009; Tavares and Eva, 2012). As challenges persist even after elaborate training efforts have been designed, researchers in health professions education have begun to consider issues directly linked to the cognitive processes of the raters (Govaerts et al., 2011; Kogan et al., 2011; Williams et al., 2003). For example, Ginsburg, et al. (2010) uncovered a dissonance between how raters think about competence and the competency based frameworks that are often used to define and assess trainees and health professionals. Consistent with these observations, Crossley et al. (2011) have reported greater reliability when rater-based assessment tools are tailored to the way in which raters think of competence. Such findings have led to an emphasis being placed on the need to further understand the cognitive processes raters engage in when making assessment decisions (Eva and Hodges, 2012; Hodges, 2013). If the problems observed result from misalignment between rating task requirements and limitations associated with the human cognitive system (Tavares and Eva, 2012) it is the tasks themselves that must change rather than attempting to alter the raters through training or other efforts.

In general, health professional education assessment practices have been built in a way that suggests an under-appreciation of how difficult a cognitive challenge rating tasks create. While candidates perform in response to clinical stimuli, raters must actively detect, select, and process relevant behaviours or events (while ignoring others), integrate performance standards/expectations, avoid contamination and cognitive biases, decipher the relevance and impact of contextual influences etc. In current competency-based frameworks commonly used to guide practice in this domain, raters have to do this not just for one dimension of performance, but for many simultaneously. For example, the commonly used mini-CEX requires raters to consider six distinct, yet inter-related, dimensions (Norcini et al., 2003).

A number of theories have emerged predicting that resource limitations associated with information processing (e.g., perception, attention, working memory and executive control) will affect rater performance when resource conflict exists (i.e., when rating demands exceed cognitive resources) (Baddeley, 1992; Navon and Gopher, 1979; Salvucci and Taatgen, 2008; Wickens, 2002). Common amongst these theories are predictions that performance can become impaired when information processing resources become overwhelmed (Baddeley, 1992; Holmboe, 2004; Lavie, 1995; Lord and Maher, 1990; Mack, 1998; Wickens and Carswell, 2006). Thinking of rating tasks in this way raises the question of whether raters' cognitive capacity may be exceeded by the cognitive demands imposed by rating tasks, thereby threatening rating quality

(DeNisi, 1996; LaMantia et al., 1999; Noel et al., 1992). For example, in real world settings, clinicians routinely maintain ultimate responsibility for patient care while assessing trainees while also having to determine the extent to which contextual variables such as the acts of other individuals might result in candidates' performance not being reflective of their individual ability. In more artificial settings, examiners must often monitor the progress of a simulation, again potentially distracting them from the task of candidate observation and judgment. In fact, the very effort that has gone into broadening medical educators' conceptions of competence by emphasizing multiple dimensions of performance may bring with it the unintended consequence of overwhelming raters when instruments require individuals to attend to all dimensions simultaneously.

Tamblyn et al., and Vu et al. have both demonstrated this sort of effect in studies of standardized patients' (SPs') accuracy in completing checklists of clinician performance. (Tamblyn et al., 1991; Vu et al., 1992) In both instances, as the number of items on a checklist increased, SP accuracy declined in a manner consistent with the suggestion that rating demands reduce rater accuracy (Tamblyn et al., 1991; Vu et al., 1992). To our knowledge, however, no one has questioned whether or not subjective performance ratings (which are not as dependent on observing specific behaviours) are similarly impacted. Williams et al. (2005) administered a 3-item scale to enable surgeons to evaluate resident performance and did not find a difference in the reliability observed when the data were analyzed when taking into account all 3 items versus when only a single item was analyzed. While this is the closest test of the hypotheses outlined in this paper, the fact that number of items was not experimentally manipulated precludes us from drawing conclusions from this study regarding the impact of rating demands on rater performance). Any detrimental cognitive impact experienced from having to evaluate 3 dimensions of performance would have exerted its influence when the ratings were collected and cannot, therefore, be corrected for at the time of analysis.

In the current study we directly and experimentally explored the influence of altering rating demands on the quality of rater-based assessments in health professions education. Based on the theories outlined above, we hypothesized that increasing the number of dimensions to be rated would lower rating quality by affecting the number of dimension relevant behaviours reported, thereby resulting in poorer ability to discriminate effectively between levels of performance and lessened inter-rater reliability.

Methods

Overview

Participants were asked to rate three pre-recorded unscripted clinical encounters of variable levels of performance (high, medium, and low) after being randomly assigned to one of four experimental conditions in a 2x2 factorial design. These conditions were intended to manipulate two ways in which rating demands are placed on raters. The first factor was the number of dimensions to be rated: Participants were assigned to rate performance using either an existing seven dimension (7D) Global Rating Scale (GRS) or a two dimension (2D) version of the same GRS created simply by removing mention of the additional 5 dimensions. The second factor involved the presence or absence of an additional extraneous, but ecologically valid, task. Half of all participants were given explicit instruction to attend to the patient's status and to the activity of other individuals observable in the video. In sum, participants were required to rate

clinical performances using either a 7 or 2 dimension GRS, with or without additional tasks simultaneously imposed upon them.

Participants

Many health professions programs include an expectation of weekly instructor-guided and simulation-based clinical practice with feedback, independent simulation-based practice, and peer observation with feedback. As a result, senior students are accustomed to observing and evaluating the clinical performance of others for the purpose of formative assessment. For this study, students were recruited using convenience sampling from three different institutions in Ontario, Canada who were sufficiently senior to have requisite clinical knowledge and skill to evaluate the performances. Despite this requisite knowledge, we accept that students are still not content experts and that this may affect our results. However, our intention in using students was first to allow us to test our hypothesis on a theoretical level, but also to explore the role of content expertise as a mitigation strategy in subsequent studies. No rater training was provided to avoid biasing the participants in a manner that might inflate or deflate the effect of the experimental intervention because this study is meant to evaluate rater-based assessment under common and natural conditions (Pelgrim et al., 2011). Furthermore, these decisions were thought appropriate given that our intention was to compare performance across conditions rather that evaluate the absolute level of performance. Participation in this study was voluntary.

Materials and Manipulations

The stimuli that were rated by all participants were three pre-recorded unscripted clinical performances. They were selected from a bank of videos recorded during a previous study that included 3 groups of candidates with different levels of training and experience (Tavares et al., 2012b). These included novices (incomplete training), entry-level clinicians (completed training but had not worked independently) and experts (clinicians with approximately 5 years of clinical practice). We randomly selected a video from each grouping for this study to create meaningful distinctions between stimuli.

The case involved attending to a cardiac patient in the presence of two first responders (i.e., other lower trained health professionals) who were portrayed by two research assistants. The patient (a mannequin) had a decreased level of consciousness secondary to a lethal arrhythmia and eventually deteriorates regardless of interventions to cardiac arrest. Each candidate was expected to assess and manage the patient as per basic cardiac life support protocol (including airway management and defibrillation) while involving the first responders as appropriate. The cases were standardized as per OSCE protocol such that each video became nine minutes in length regardless of candidate performance. Study participants were instructed that the candidates in the performances were participants in an OSCE process and that decisions were to be made regarding their ability to enter independent practice.

Factor A – Number of Dimensions to be Rated

For half of all participants, the GRS used in this study was an existing 7D GRS designed for the assessment of paramedic clinical competence. It included items corresponding to: (a) situation awareness, (b) history gathering, (c) patient assessment, (d) decision making, (e) resource utilization, (f) communication, and (g) procedural skill (Tavares et al., 2012b). This scale was

rigorously designed using task analysis, a Delphi procedure, and psychometric testing to achieve construct representation (Streiner and Norman, 2008). Scores were collected for each dimension using a 7-point adjectival scale with anchors ranging from 1 (unsafe and unsuitable for supervised practice) to 7 (highly competent, ready for independent practice, a model for others).

For the other half of participants, we attempted to reduce rating demands by decreasing the number of dimensions participants were asked to consider from seven to two. We chose to use two dimensions because literature-based analyses suggest that raters' judgments tend to collapse into two dimensions even when they are asked to consider a larger number (Lurie et al., 2009; Williams et al., 2003). To create the 2D scale, we simply deleted 5 of the items from the existing GRS. In selecting the 2 dimensions to be included, we considered a combination of both conceptual and empirical distinctiveness as determined during the scale's initial development and critical appraisal process (Tavares et al., 2012b). We eventually selected "History Gathering" and "Procedural Skill" (inter-item correlation of .66). Other pairs of dimensions (e.g., History Gathering and Communication) also demonstrated evidence of empirical distinctiveness (i.e., inter-item correlation of .63) but conceptually seemed more related than the pairing we selected. Selecting dimensions that were relatively uncorrelated allowed us to prompt the consideration of distinct aspects of performance.

Factor B – Presence or Absence of Additional Ecologically Valid Tasks

Half of all participants in both the 7D and 2D groups were explicitly instructed to monitor both the first responders' behaviours and the patient's condition (in addition to the candidates' behaviours) to identify any need for additional intervention. While rating candidate performance, raters often need to simultaneously monitor the patient's condition (e.g., seeking additional assessment choices, considering degree of patient severity and need for/urgency of interventions) and attend to the context in which the case is occurring (e.g., the behaviour or actions of other team members, the role of environmental cues). Participants given this additional task were provided with the following instructions:

> "*As you observe the candidate, we would like you to also monitor the patient's condition, and consider your own management of this patient. At the end of the video, we will ask you to identify and verbalize assessment procedures and/or interventions that the patient in this case is in need of and to identify which of the two transfer company employees [i.e., standardized actors in the case] is responsible for the situation they are in and why, so please monitor their discussion as well.*"

Procedure

Following consent to participate and completion of a demographic questionnaire, participants were randomly assigned to one of the four experimental conditions using a computer generated random list. A brief orientation to the GRS and study procedure was provided. First, participants were oriented to the dimensions included on the assigned GRS, the rating labels and the associated definitions for each. Participants were informed that other dimensions of performance may be observable, but that we were interested only in the dimensions included on their assigned scale.

Participants were instructed to verbally identify any behaviours or events that were directly relevant to the dimensions on the rating scales assigned while watching the video. They did so using a laptop computer with headset and wireless microphone. For each identified behaviour, participants were asked to specifically indicate three features: (a) the specific behaviour detected; (b) the dimension of relevance to the behaviour; and, (c) whether the behaviour contributed to generating a sense of competence (by stating the behaviour was positive) or incompetence (by stating the behaviour was negative) for the candidate being observed. For example, participants might state "he repositioned the airway, that's a procedural skill and it is positive". Participants were given the freedom to assign behaviours to more than one dimension if they felt that a particular behaviour applied to more than one. This method of using verbal reports as data is described in detail elsewhere (Chi, 1997) and was intended to provide an indication of which behaviours or aspects of performance were explicitly noted by raters. All participants, regardless of condition, were required to think aloud in the same manner.

As a measure of cognitive load, we used a commonly applied strategy of requiring participants to perform a secondary task concurrently with the primary task of assessing candidate performance. We applied a vibrotactile device to each participant's left forearm. At random intervals (from 10 – 90 s) during the video, a signal was sent to the device using Bluetooth technology causing it to vibrate (Byrne et al., 2010). Participants were instructed to focus on the primary task (i.e., rating candidate performance), but to terminate the vibration as quickly as possible, when it was detected, by depressing a button on the device. Performance on the secondary task (i.e., response time) was intended to reflect the level of cognitive demand associated with the primary task in that the more challenged the participant is by the primary task the longer it should take to respond to the secondary task (Paas et al., 2003). Participants were given an opportunity to feel the stimulus and practice the procedure before commencing the study. In addition, following each video, participants were asked to rate the candidate's performance using the assigned GRS and to complete a subjective self-report measure, known as the NASA-Task Load Index, to indicate their perceived mental burden (Hart, 2006; Hart and Staveland, 1988). The Index includes 6 subscales (mental, physical and temporal demands, frustration, effort and performance) the combination of which is intended to represent the subjective mental (and physical) workload associated with a task (Hart, 2006).

Outcome Measures and Analysis

Psychometrically, assessments are deemed to be sound, in part, when they are based on accurate observation, demonstrate an ability to discriminate between distinct levels of performance and demonstrate consistency in ratings across individuals (Streiner and Norman, 2008). As such, we used three outcome measures:

(1) Number of dimension relevant behaviours identified: To quantify observational acuity we began by transcribing verbatim audio recordings and capturing all behaviours or events identified across participants in all groups for each video. To be included under a particular dimension, the identified behaviour or event must have been clearly stated by the participant in the manner described above and aligned with the definition included on the rating scale. Behaviours or events identified by 2 or more participants were included as part of the criterion. Across all think aloud responses received from all participants we counted the total number of distinct behaviours or events participants identified for each of the 2 dimensions common to all groups (history gathering and procedural skill). We

then calculated, for each participant, the proportion of possible behaviours identified using the criterion described above. This participant-driven process eliminated the need for interpretation or ambiguity in the data. However, to investigate the robustness of the results, we also searched the transcriptions for responses that were included in our criterion list but were placed in another dimension by participants and we re-analyzed the data using this more comprehensive coding strategy. Performance on this measure was compared across conditions using ANOVA.

(2) Discrimination of level of performance: The scores assigned to the two dimensions common to both groups were used to determine participants' ability to discriminate between levels of performance by treating the high, medium, and low performance videos as levels of a within subjects variable in a repeated measures ANOVA. This analysis was done for each condition independently and post-hoc t-tests were used to compare the mean scores across video in a pairwise manner.

(3) Reliability: Generalizability theory was used to calculate the reliability of the scores assigned to the 2 dimensions common to all groups (Brennan, 2001). Through ANOVA, the variance in scores can be parceled out into variance attributable to the candidates displayed in the videos, the raters who served as participants, the items included on the scale, and the various interactions between these variables. Generalizability theory then allows one to determine what proportion of variance can be attributed to the subject of measurement (video), as is desirable whenever one wishes to make claims about the reliability of test scores. In this way we could calculate internal consistency (i.e., the extent to which scores on one item are associated with scores on another item), and inter-rater reliability (i.e., the extent to which raters consistently differentiated between the candidates presented on video). Finally, the amount of error attributable to raters could also be identified.

We hypothesized that being asked to attend to seven dimensions and the instruction to complete additional tasks would decrease observational comprehensiveness, impair ability to discriminate between groups, and decrease reliability relative to being asked to attend to only two dimensions and the lack of additional tasks. Significance level was set at .05 for all ANOVA analyses. SPSS Ver. 19 and G-String Ver. IV were used where appropriate.

Results

A total of 44 raters were recruited to participate in this study. Seventy-five percent were male and 25% percent were female. The majority of participants listed a university degree (n=16) as the highest level of education completed, followed by high school (n=15), community college (n=12) and graduate level education (n=1). All indicated experience rating clinical performance as senior students, but none of the participants had received any formal rater training.

Number of relevant behaviours noted

Participants collectively observed an average (across video) of 18 and 28 distinct behaviours or events reflective of the "history gathering" and "procedural skills" dimensions, respectively. Using this as the criterion, the mean proportion of identified possible behaviours was calculated for each individual (see Table 1) and compared across conditions.

A 2 (7D vs. 2D) x 2 (Additional task vs. No additional task) x 3 (Video) ANOVA was performed on the proportion of behaviours identified for each dimension (history gathering and procedural skills) using SPSS with dimension and task treated as between subjects factors and video as a within subject factor. In both instances a main effect of 2D vs. 7D was observed (History Gathering $F(1,33)=17.5$, $p<0.001$, $d=1.27$; Procedural skills $F(1,33)=18.2$, $p<0.001$, $d=.96$) with more distinct behaviours being reported as relevant in the 2D condition relative to the 7D condition. To investigate the robustness of our initial results, we re-analyzed the data by including noted behaviours that fit the 2 dimensions of interest, but were labeled by participants as indicating one of the other 5 dimensions. In this way, we gave credit for detection/selection, irrespective of the dimension to which the observation was assigned by participants. The proportions increased slightly, but the analysis yielded identical conclusions: Participants in the 2D condition showed greater observational acuity than those in the 7D condition (History Gathering $F(1,33)=16.0$, $p<0.001$; Procedural skills $F(1,33)=9.8$, $p<0.01$).

The presence or absence of the additional task intended to distract raters had no influence on observational acuity, nor did it interact with video or the number of dimensions. As such, this task did not appear to have influenced raters and was not considered further.

Ability to Discriminate Between Levels of Performance

Using the GRS scores assigned by raters on each of the 2 dimensions common to both groups as the dependent variables, a 2 (7D vs. 2D) x 3 (Video) ANOVA with dimensions as the between subject factor and video as the within subject factor was performed to determine if experimental condition influenced ability to discriminate between levels of performance. For the dimension History Gathering, a main effect of Video was observed, indicating that individuals were able to differentiate between levels of performance $F(2,76) = 52.3$, $p = < .001$. In addition, a significant Video x Dimension interaction was observed $F(2,76) = 3.5$, $p = .04$ indicating that the differences in ratings across video were greater in the 2D group relative to the 7D group as illustrated in Table 2. For the dimension Procedural Skill, similar results were observed, with a main effect of Video $F(2,76) = 84.3$, $p = <.001$, and Video x Dimension interaction $F(2,76) = 4.5$, $p = .01$. In both instances, post hoc analyses revealed that the mean scores assigned in the 7D experimental condition reached significance on 2 of 3 pairwise comparisons whereas they reached significance for all 3 pairwise comparisons in the 2D experimental condition.

Reliability

Using scores assigned by raters on the 2 dimensions common to all groups, Generalizability theory was used to determine how much variance could be attributed to each variable included in the study design and reliability was calculated for both the 7D and 2D conditions independently. Individual variance components, the percentage of total variance attributable to each facet and their interactions are reported in Table 3 along with the formulas used to calculate the g-coefficients and d-study results. The variance attributable to the facet of differentiation (i.e., video) for the 7D condition was 44.7% compared to 67.7% in the 2D condition, suggesting that raters in the 2D condition were better able to discriminate between videos. The inverse of this is that variance attributable to the main effect of rater and the "rater-x-video interaction (psychometrically referred to as rater-related measurement error) was 6.45% and 25.35%, respectively in the 7D condition, compared to 2.22% and 8.86%, respectively, in the 2D condition, indicating that rater error increases with higher rating demands.

Converting these numbers to reliability coefficients revealed that inter-rater reliability improved as the number of dimensions decreased from 7 (G=.45) to 2 (G=.70). The internal consistency did not change substantially as a function of number of dimensions evaluated (7D G=.70, 2D G=.77). This indicates that the ability to consistently differentiate between candidates improved as rating demands decreased.

Assessment of Load using Secondary Tasks

A one-way ANOVA was performed using SPSS on both NASA-TLX scores and the vibrotactile tasks response time outcomes and revealed no significant effects. Mean scores on the NASA-TLX in the 7D and 2D conditions, respectively, were 50.8 (SD=2.4) and 51.4 (SD=2.4), ($F(1,39)=1.2$, $p=.28$). Mean response times on the secondary task in the 7D and 2D conditions, respectively, were 874.0 (SD=119.2) msecs and 731.9 (SD=119.2) msecs ($F(1,39)=0.7$, $p=.40$).

Discussion

This experimental study explored the impact of increasing rating task demands on various indicators of rater performance. The results provide support for the hypothesis that as the number of dimensions to be simultaneously rated increases, the number of relevant observations, the ability to discriminate between levels of performance and inter-rater reliability all decline. As a result, this study contributes evidence suggestive of a mechanism (i.e., mental workload) through which rater-based assessments that require raters to simultaneously assess many factors can lead to rater idiosyncrasies.

Consistent with modern theories of working memory (Wickens, 2002, 2008), asking raters to consider multiple dimensions simultaneously should result in greater opportunity for dimension specific behaviours or events to be idiosyncratically missed or variably interpreted, leading to impaired rater performance. As a result, this study is consistent with the work of Byrne et al., (2010) and others in medical education and elsewhere that has linked performance impairments to cognitive demands (Alexander, 2000; Byrne et al., 2010; Cantin et al., 2009; Tsang and Vidulich, 2006; Yurko et al., 2010; Zheng et al., 2010).

That said, suggesting mental workload as the mechanism that led to our findings is necessarily speculative at this point. It was expected that the secondary tasks used (i.e., vibrotactile stimulation and prompting individuals to attend to patient related information and monitor bystander behaviour) and the NASA-TLX scores would yield a further test of working memory impedance. The lack of effect on these variables suggests we were either unable to induce or capture changes in working memory, or that task switching was occurring without inducing load. Prompting raters to attend to patient related factors may have had no effect given the automaticity with which individuals with clinical background are likely to attend to such factors even when not prompted to do so.

Alternatively, the 7D vs. 2D effects observed in the context of no apparent effect on working memory might suggest that those in the 7D condition were impaired due to the scale creating a greater degree of misalignment with raters' natural cognition. That is, perhaps being pushed to think about all 7 dimensions took individuals further from the way in which they normally conceive of trainees' competence in a manner consistent with Crossley et al.'s (2011) argument rather than being attributable to working load limitations per se.

Despite the lack of effect on the second experimental factor, the difference observed between the 7D and 2D conditions suggests that considering the alignment (or lack thereof) between the demands associated with a rating task and inherent cognitive capacity can inform challenges commonly observed in rater-based assessment. For instance, Lamantia et al., (1999) had faculty evaluate residents' clinical skills using both checklists and global rating scales during standardized patient encounters and noted significant variation in what faculty observed. The authors attributed the lack of agreement to variation in skill. Yeates et al., (2012) used a think aloud procedure while raters observed scripted video-based clinical performances and also noted significant variation in what raters considered. The authors in this study attributed the results to differential salience, criterion uncertainty and differences in how the collected information was integrated by examiners. The results of the current study suggest an explanation (i.e., mental workload) that may provide a causal mechanism for these differences and indicates why raters who observe the same performance may nonetheless generate variable impressions.

Under higher task demands raters may engage mental shortcuts or strategies through the application of schemas, heuristics, or the adoption of solutions that 'satisfy' rather than 'optimize' the rater's needs (Tavares and Eva, 2012). For example, if asked to assess many dimensions simultaneously, raters may engage in either degraded concurrent processing and/or serial processing of available information, both of which may be problematic (Wickens and Carswell, 2006). Degraded concurrent processing suggests raters consider multiple dimensions concurrently, but that those considerations suffer relative to considering each dimension in isolation. Here, raters attempt to satisfy all the demands associated with the rating task, but may unintentionally (and idiosyncratically) fail to do so. Serial processing suggests only one (or few) dimensions may be considered optimally at one time while others are negatively affected due to sequential constraints. Here, raters may make idiosyncratic choices about which dimensions or behaviours to divert their attention to at a cost to others. Both possibilities are likely to contribute to rater variation, but would require further study to tease apart.

Cognitive strategy notwithstanding, for assessment developers this study suggests that attention toward the demands imposed on raters may need to be considered in scale or assessment strategy development in order to maximize rater based assessments. For instance, the global rating scale used in this study (i.e., the original 7D version) followed many conventions in its development process (Streiner and Norman, 2008; Tavares et al., 2012b) and closely matches the development of similar clinical performance scales (Fletcher et al., 2003; Kim et al., 2006; Norcini et al., 1995). Developmental decisions, such as ensuring adequate construct representation, are often made in the absence of consideration of the impact on raters yet inherent rater ability to process presented information might have a fundamental influence on the utility of the assessment practice.

Limitations

The failed results on the NASA-TLX and the lack of effect of additional tasks (factor B / extraneous task) make it difficult to conclude the results observed were directly linked to general mental workload rather than being specific to the number of dimensions to be rated. The cognitive demands associated with our additional tasks may have been integrated into all conditions due to their aligning with what raters would naturally do in these situations. The second limitation is the recruitment of students as raters. The differences that exist between novice and expert raters (e.g., knowledge base / domain knowledge) may have had an impact on

the outcomes of this study. Future research should evaluate whether the development of clinical expertise, rating expertise, or rater training mitigate the results we observed. Third, our behaviour observation data only reflects omission errors and not commission errors. However, our study objective was to evaluate what the participants could "see" across experimental conditions and, hence, we do not believe this to invalidate the results. Finally, the act of verbalizing and the need to complete the secondary vibrotactile task throughout the assessment process may have contributed additional cognitive load and may, therefore, have affected performance overall. This methodology was consistently applied across experimental condition and, as such, we do not think it can account for the differences observed.

Conclusions

For educators and researchers, the results of this study contribute to the efforts aimed at understanding rater based assessments of clinical competence by uncovering a potential mechanism associated with rating quality and/or rater variance. It appears that as the demands imposed on raters increases (in this case by increasing the number of dimensions to be rated simultaneously), it negatively affects raters' observations, reliability and their ability to discriminate between dimensions. A review of the assessment literature would reveal that this issue has largely remained unknown or ignored given the fact that many assessments tools and processes are designed without any consideration for the impact design choices have on rating demands or rater performance. In extending this research, some of the challenges the assessment community faces will be (a) identifying predictors (e.g., contextual influences) of when rating demands and cognitive resources become poorly aligned, and (b) determining effective methods of measuring mental workload or cognitive alignment in clinical assessment contexts. Determining what process, design, or other strategies (e.g. using mental aids, attempts to improve alignment with rater conceptions of performance) can be used to address this issue or how the assessment community resolves the apparent contradiction between ensuring comprehensive construct representation and mental workload requirements will require additional research. A promising area of assessment research will be to investigate and recommend design changes when conditions of multi-task resource overload might exist.

Acknowledgments:

The authors would like to thank the administration, faculty and students of Centennial College, Georgian College and Humber College paramedic programs for their support in this study. The authors would also like to thank Ellen Ironside and the members of the University of Toronto Wilson Centre for Health Professions Education for their support.

Funding:

This study was generously supported by the Health Sciences Department and Applied Research and Innovation Centre at Centennial College in Toronto, On., Canada.

Other Disclosures:

None.

Ethical Approval

Ethical approval was provided by Centennial College (REB#124), Humber College (REB # 0164) and Georgian College Research Ethics Boards (no REB. #).

Disclaimer:
None.

Previous Presentations:
University of Toronto Wilson Centre Research Day.

Table 1: Observational acuity: Mean proportion (and 95% Confidence Intervals) of possible behaviours identified as a function of Dimension (History gathering vs. Procedural skills), the Number of Dimensions evaluated (7D vs. 2D), and the Additional task variable (Present vs. Absent)

	History Gathering			Procedure skills		
	Additional task			Additional Task		
Dimensions	Present	Absent	Mean	Present	Absent	Mean
7D	17.8 (9.4-26.2)	14.4 (6.0-22.8)	16.1 (10.1-22.0)	10.9 (5.5-16.3)	10.1 (4.7-15.5)	10.5 (6.7-14.4)
2D	30.9 (22.5-39.3)	35.4 (27.4-43.4)	33.2 (27.4-39.0)	19.5 (14.0-24.9)	24.1 (18.9-29.2)	21.8 (18.0-25.5)
Mean	24.3 (18.4-30.3)	24.9 (19.1-30.7)		15.2 (11.4-19.0)	17.1 (13.3-20.8)	

-

Table 2a: Mean GRS scores (and 95% Confidence Intervals) assigned to each video for the History Gathering dimension as a function of number of dimensions assessed.

History Gathering				
Dimensions	High performance	Median performance	Low performance	Comparisons
7D	4.71 (4.06-5.37)	4.07 (3.41-4.73)	2.79 (2.22-3.35)	High Vs. Med.: p = .27 High Vs. Low: p= < .01 Med. Vs. Low: p= < .01
2D	5.69 (5.26-6.11)	4.75 (4.25-5.25)	3.25 (2.65-3.85)	High Vs. Med.: p= < .05 High Vs. Low: p= < .01 Med. Vs. Low: p= < .01

Table 2b: Mean GRS scores (and 95% Confidence Intervals) assigned to each video for the Procedural Skill dimension as a function of number of dimensions assessed.

Procedural Skill				
Dimensions	High performance	Median performance	Low performance	Comparisons
7D	4.64 (4.16-5.13)	4.43 (3.84-5.02)	2.71 (1.85-3.57)	High Vs. Med.: p = .86 High Vs. Low: p= < .01 Med. Vs. Low: p= < .01
2D	5.63 (5.05-6.20)	4.75 (4.39-5.11)	2.38 (1.90-2.85)	High Vs. Med.: p= < .05 High Vs. Low: p= < .01 Med. Vs. Low: p= < .01

Table 3: Generalizability results and variance attributable to facets included in the study as a function of experimental condition

	7D		2D	
Effect	**Variance**	**% of Total Var.**	**Variance**	**% of Total Var.**
Video	0.97	44.70	2.14	67.72
Rater	0.14	6.45	0.07	2.22
Items	0.00	0.00	0.00	0.00
Video x Rater	0.55	25.35	0.28	8.86
Video x Items	0.00	0.00	0.06	1.89
Rater x Items	0.00	0.00	0.07	2.22
Video x Rater x Items	0.51	23.50	0.54	17.11
Inter rater reliability:	G = 0.45; G_n = .91		G = 0.70; G_n = .96	
Internal consistency:	G = 0.70; G_n = .96		G = 0.77; G_n = .97	

7D = groups assigned seven dimensions; 2D = groups assigned two dimensions; Var. = Variance component.

G Formulae:

Inter-rater Reliability: [Var(v) + Var(vi)] / [Var(v) + Var(r)+Var(i)+Var(vr) + Var(vi) + Var(ri) + Var(vri)]

Internal Consistency: [Var(v) + Var(vr)] / [Var(v) + Var(r)+Var(i)+Var(vr) + Var(vi) + Var(ri) + Var(vri)]

G_n Formulae:

Inter-rater Reliability: [Var(v) + Var(vi)/2] / [Var(v) + Var(r)/10+Var(i)/2+Var(vr)/10 + Var(vi)/2 + Var(ri)/20 + Var(vri)/20]

Internal Consistency: [Var(v) + Var(vr)/10] / [Var(v) + Var(r)/10+Var(i)/2+Var(vr)/10 + Var(vi)/2 + Var(ri)/20 + Var(vri)/20]

References:

Alexander, A.L. (2000*).* Examining the relationship between mental workload and situation awareness in a simulated air combat task. DTIC Document.

Brennan, R.L. (2001). *Generalizability theory*. New York NY. USA. Springer Verlag.

Byrne, AJ, Oliver, M., Bodger, O., Barnett, WA, Williams, D., Jones, H., & Murphy, A. (2010). Novel method of measuring the mental workload of anaesthetists during clinical practice. *British journal of anaesthesia, 105*(6), 767-771.

Cantin, V., Lavallière, M., Simoneau, M., & Teasdale, N. (2009). Mental workload when driving in a simulator: Effects of age and driving complexity. *Accident Analysis & Prevention, 41*(4), 763-771.

Chi, M. (1997). Quantifying qualitative analyses of verbal data: A practical guide. *Journal of the Learning Sciences, 6*(3), 271-315.

Crossley, J., Johnson, G., Booth, J., & Wade, W. (2011). Good questions, good answers: construct alignment improves the performance of workplace based assessment scales. *Medical Education, 45*(6), 560-569.

DeNisi, A.S. (1996). *A cognitive approach to performance appraisal: A program of research*: New York NY. USA. Routledge.

Downing, S.M. (2004). Reliability: on the reproducibility of assessment data. *Medical Education, 38*(9), 1006-1012.

Downing, S.M. (2005). Threats to the validity of clinical teaching assessments: What about rater error? *Medical Education, 39*(4), 353-355.

Eva, K.W., & Hodges, B.D. (2012). Scylla or Charybdis? Can we navigate between objectification and judgement in assessment? *Medical Education, 46*(9), 914-919.

Fletcher, G., Flin, R., McGeorge, P., Glavin, R,. Maran, N., & Patey, R.. (2003). Anaesthetists' Non-Technical Skills (ANTS): evaluation of a behavioural marker system. *British Journal of Anaesthesia, 90*(5), 580-588.

Ginsburg, S., McIlroy, J., Oulanova, O., Eva, K. W., & Regehr, G. (2010). Toward Authentic Clinical Evaluation: Pitfalls in the Pursuit of Competency. *Academic Medicine, 85*(5), 780-786.

Govaerts, M., van de Wiel, M., Schuwirth, L., van der Vleuten, C., & Muijtjens, A. (2013). Workplace-based Assessment: rater's performance theories and constructs. *Advances in Health Sciences Education, 18*(3), 375-396.

Green, M.L., & Holmboe, E. (2010). Perspective: The ACGME Toolbox: Half Empty or Half Full? *Academic Medicine, 85*(5), 787-790.

Haber, R.J., & Avins, A.L. (1994). Do ratings on the American Board of Internal Medicine Resident Evaluation Form detect differences in clinical competence? *Journal of General Internal Medicine, 9*(3), 140-145.

Hart, S.G. (2006). NASA-task load index (NASA-TLX); 20 years later. *In Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting* (pp. 904-908). Santa Monica CA: Human Factors & Ergonomics.

Hart, S.G., & Staveland, L.E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Human Mental Workload,* (pp. 139–183). Amsterdam. Elsevier.

Hodges, B. (2013). Assessment in the post-psychometric era: Learning to love the subjective and collective. *Medical Teacher, 35*(7), 564-568.

Holmboe, E.S. (2004). Faculty and the observation of trainees' clinical skills: problems and opportunities. *Academic Medicine, 79*(1), 16-22.

Huwendiek, S., Mennin, S., Dern, P., Friedman Ben-David, M., Van Der Fleuten, C., Tonshoff, B., Nikendei, C. (2010). Expertise, needs and challanges of medical educators: Results of an international web survey. *Medical Teacher, 32*(11), 912-918.

Kim, J, Neilipovitz, D, Cardinal, P, Chiu, M, & Clinch, J. (2006). A pilot study using high-fidelity simulation to formally evaluate performance in the resuscitation of critically ill patients: The University of Ottawa Critical Care Medicine, High-Fidelity Simulation, and Crisis Resource Management I Study. *Critical care medicine, 34*(8), 2167-2174.

Kogan, J.R., Conforti, L., Bernabeo, E., Iobst, W., & Holmboe, E. (2011). Opening the black box of clinical skills assessment via observation: a conceptual model. *Medical education, 45*(10), 1048-1060.

LaMantia, J, Rennie, W, Risucci, DA, Cydulka, R, Spillane, L, Graff, L., & Kleinschmidt, K. (1999). Interobserver Variability among Faculty in Evaluations of Residents Clinical Skills. *Academic Emergency Medicine, 6*(1), 38-44.

Lavie, N. (1995). Perceptual load as a necessary condition for selective attention. *Journal of Experimental Psychology: Human Perception and Performance, 21*(3), 451-468.

Lord, R.G., & Maher, K.J. (1990). Alternative information-processing models and their implications for theory, research, and practice. *The Academy of Management Review, 15*(1), 9-28.

Lurie, S.J, Mooney, C.J., & Lyness, J.M. (2009). Measurement of the general competencies of the Accreditation Council for Graduate Medical Education: a systematic review. *Academic Medicine, 84*(3), 301-309.

Mack, A., & Rock, I. (1998). Inattentional Blindness. Cambridge, MA: MIT Press.

Navon, D., & Gopher, D. (1979). On the economy of the human-processing system. *Psychological review, 86*(3), 214-255.

Noel, G.L., Herbers, J.E., Caplow, M.P., Cooper, G.S., Pangaro, L.N., & Harvey, J. (1992). How well do internal medicine faculty members evaluate the clinical skills of residents? *Annals of Internal Medicine, 117*(9), 757-765.

Norcini, J., Blank, L., Arnold, G., & Kimball, H. (1995). The mini-CEX (clinical evaluation exercise): A preliminary investigation. *Annals of Internal Medicine, 123*(10), 795-799.

Norcini, J., Blank, L., Duffy, F., & Fortna, G. (2003). The mini-CEX: a method for assessing clinical skills. *Annals of Internal Medicine, 138*(6), 476-481.

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

Pelgrim, E., Kramer, A., Mokkink, H., Van den Elsen, L., Grol, R., & Van der Vleuten, C. (2011). In-training assessment using direct observation of single-patient encounters: a literature review. *Advances in Health Sciences Education, 16*(1), 131-142.

Salvucci, D.D., & Taatgen, N.A. (2008). Threaded cognition: An integrated theory of concurrent multitasking. *Psychological Review, 115*(1), 101-130.

Streiner, D.L, & Norman, G.R. (2008). *Health measurement scales: a practical guide to their development and use* (Fouth ed.). New York: Oxford Univeristy Press.

Tamblyn, R.M., Klass, D.J., Gail, K., Schnabl, K., & Kopelow, M.L. (1991). Sources of Unreliability and Bias in Standardized-Patient Rating. *Teaching and Learning in Medicine, 3*(2), 74-85.

Tavares, W., & Eva, K.W. (2012a). Exploring the Impact of Mental Workload on Rater-Based Assessments. *Advances in Health Sciences Education,18(2), 291-303.*

Tavares, W., Boet, S., Theriault, R., Mallette, T., & Eva, K.W. (2012b). Global Rating Scale for the Assessment of Paramedic Clinical Competence. *Prehospital Emergency Care, 17*(1), 57-67.

Thompson, W. G., Lipkin, M., Jr., Gilbert, D. A., Guzzo, R. A., & Roberson, L. (1990). Evaluating evaluation: assessment of the American Board of Internal Medicine Resident Evaluation Form. *Journal of General Internal Medicine, 5*(3), 214-217.

Tsang, P.S., & Vidulich, M.A. (2006). *Mental Workload and Situation Awareness*. Hoboken, NJ: Wiley.

Vu, Nu Viet., Marcy, M., Colliver, J.A., Verhulst, S.J., Travis, T.A., & Barrows, H.S. (1992). Standardized (simulated) patients' accuracy in recording clinical performance check-list items. *Medical Education, 26*(2), 99-104.

Wickens, C.D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science, 3*(2), 159-177.

Wickens, C.D. (2008). Multiple resources and mental workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 50*(3), 449-455.

Wickens, C.D., & Carswell, C. (2006). *Handbook of human factors and ergonomics, Third Edition*. Hoboken, NJ: Wiley

Williams, R.G., Verhulst, S., Colliver, J.A., & Dunnington, G.L. (2005). Assuring the reliability of resident performance appraisals: more items or more observations? *Surgery, 137*(2), 141-147.

Williams, RG, Klamen, DA, & McGaghie, WC. (2003). SPECIAL ARTICLE: Cognitive, Social and Environmental Sources of Bias in Clinical Performance Ratings. *Teaching and Learning in Medicine, 15*(4), 270-292.

Yeates, P., O'Neill, P., Mann, K., & Eva, K.W. (2012). Seeing the same thing differently. *Advances in Health Sciences Education*, *18*(3), 325-341.

Yurko, Y.Y., Scerbo, M.W., Prabhu, A.S., Acker, C.E., & Stefanidis, D. (2010). Higher mental workload is associated with poorer laparoscopic performance as measured by the NASA-TLX tool. *Simulation in Healthcare, 5*(5), 267-271.

Zheng, B., Cassera, M.A., Martinec, D.V., Spaun, G.O., & Swanström, L.L. (2010). Measuring mental workload during the performance of advanced laparoscopic tasks. *Surgical Endoscopy, 24*(1), 45-50.

Chapter 5 – Selecting and Simplifying: Rater Performance and Behaviour When Considering Multiple Competencies

Abstract

Introduction

The validity associated with rater judgment in assessments of clinical competence is dependent on a rater's cognitive performance. In this study we asked: what effect does broadening raters' focus by asking them to consider multiple competencies have on indicators of rating quality? Second, we explored the cognitive strategies raters engage when asked to consider multiple competencies simultaneously.

Methods

In this parallel mixed methods study, participants rated 3 recorded clinical performances within a 2x2 factorial design. Half of the participants were asked to rate seven dimensions of performance (7D) and half were asked to rate only 2 (2D). The second factor involved the requirement (or lack thereof) to rate the performance of actors participating in the simulation. We measured the identification of dimension-relevant behaviours and calculated the reliability of the scores assigned. Second, we analyzed data from semi-structured post-task interviews to explore the strategies associated with rating under conditions of broadened focus.

Results

Using the two dimensions common to both groups, ANOVA revealed that participants who were asked to rate only 2 dimensions considered more dimension relevant features than those asked to rate 7 dimensions (Procedural Skill dimension: 36.2% (95% CI=32.5-40.0) vs. 22.4% (95% CI=20.8-26.3); History Gathering dimension: 38.6% (95% CI=33.5-42.9) vs. 24.0% (95% CI=21.1-26.9); $p<.05$ in both instances). Also, Decision studies revealed that the 2D group achieved higher inter-rater reliability (G_10=.62; relative to the 7D group (G_10=.48)). The requirement to complete an additional rating task did not have an effect. Raters in the 7-dimension condition identified many sources of cognitive load and idiosyncratic cognitive strategies used to reduce load during the rating task.

Conclusions

As rating demands increase, indicators of rating quality decline. The strategies raters engage when asked to rate many dimensions simultaneously are varied and appear to yield idiosyncratic efforts to reduce cognitive effort, which may affect the degree to which raters make judgments based on comparable information.

Introduction

Ensuring accurate assessment of clinical performance remains a priority in medical education and continues to challenge the field (Huwendiek, 2010). In an effort to optimize assessment and address the breadth of competence expected of physicians, many have argued for a greater reliance on rater judgment (Eva and Hodges 2012; Hodges 2013). This has led to a body of research in which the "rater" is positioned as the object of study with focus placed upon the role of cognitive biases and expertise as it applies to rater performance, the effect of rater training, understanding social judgment and developing rater process and performance theories (Govaerts, et al. 2011; Gingerich, et al. 2011; Govaerts, et al. 2011; Kogan, et al. 2011; Yeates, et al. 2012). The influence of the demands that the tasks to be completed place on raters have seldom been considered as this factor is often overshadowed by attention to construct representation (i.e., the creation of the scale) or process considerations (i.e., how the ratings will be collected). As construct representation becomes more elaborate and raters continue to be identified as a major source of error in performance-based assessments (Downing. 2005; Downing and Yudkowsky 2009), however, it is necessary to consider factors that may contribute to rater difficulty. This study explores the role of rating task complexity and the alignment, or lack thereof, between rater cognitive capacity and rating demands.

Assessment of clinical competence is a complex cognitive task with many demands often imposed unintentionally (Govaerts, et al. 2011; Tavares and Eva 2012, Yeates, et al. 2012). As exam candidates respond to clinical challenges, rating their performance involves a process of attention (e.g., actively detecting elements of the performance), retrieval of information (e.g., evaluating the adequacy with which the performance matches a known standard), and processing (e.g., assigning appropriate weightings, ignoring irrelevant data) to ultimately reach some form of categorical decision regarding the candidate's clinical performance ability (DeNisi 1996). A number of theories in cognitive psychology suggest that many of the cognitive functions necessary for rating are capacity limited and may therefore be a source of difficulty. For instance, Baddeley's (1992) concept of working memory, described as the site for temporary storage and manipulation of information, plays an important role in the performance of cognitive tasks(Baddeley 1992). As this system is limited in capacity, there is in an ever-present risk of performance failures under conditions of high demand. Wicken's (2006) information processing theory, which includes attention, working memory, long-term memory, processing and execution, similarly makes predictions regarding performance failures when demands on the system are high (Wickens 2002, Wickens and Carswell 2006). Finally, Sweller's cognitive load theory (CLT), applied mainly to learning, defines different types of load and how they predict learning behaviour or performance in a limited capacity system (Sweller 1988). A recent review exploring these concepts as they apply to rater-based assessments suggested that these theories may equally be relevant to, predictive of, and explanatory of rater performance (Tavares and Eva 2012). A main finding of this review was that as mental workload increases, raters may be expected to engage in simplifying strategies or mental shortcuts that, when applied, can contribute to judgment error.

With this in mind, the movement in the health professions toward broadly considering various competencies along with greater emphasis on work-based assessment may be imposing considerable demands on raters (Byrne, et al. 2014). For instance, recent research found that as

the number of dimensions of performance to be rated increased, identification of dimension relevant behaviours and discrimination between levels of performance became impaired (Tavares and Eva (submitted)). That is, requiring raters to consider multiple complex dimensions such as communication skill and situation awareness simultaneously without considering the cognitive demands that doing so might impose, may result in a lack of alignment between imposed rating demands and inherent cognitive capacity. Raters may have difficulty considering all relevant behaviours in a clinical performance(Holmboe 2004), leading to idiosyncratic variations in what is perceived(Herbers, et al. 1989) and, in turn, differences in processing and weighting of available information and, ultimately, differences in judgments (LaMantia, et al. 1999).

Our previous tests of these hypotheses were performed using relatively novice raters as participants. (Tavares and Eva (submitted)) As content expertise develops and/or rating practice matures, efficiencies in cognitive processing may be gained through related schema acquisition and automation, ultimately increasing resiliency to various rating demands. The purpose of this study was to extend the earlier work of Tavares and Eva by examining the alignment, or lack thereof, between cognitive demands imposed by a rating task and the human processing capabilities of raters who are known to have considerable content knowledge and rating experience. This study further extends previous work by exploring raters' understanding of how they manage the workload imposed on them when asked to rate clinical performance under high demand conditions. Our fundamental research question was: In assessing clinical competence, what effect does broadening raters' focus by requiring them to consider multiple dimensions of performance have on indicators of rating quality (e.g., identification of relevant behaviours and inter-rater reliability)? This issue is of theoretical and practical importance as it is likely to inform assessment design, scale development and rater training strategies aimed at improving rater-based clinical performance assessments.

Methods

Study Overview

We used a parallel, mixed methods study design to collect quantitative and qualitative data simultaneously while analyzing the data of each strand concurrently but independently (Creswell and Clark 2007).

Faculty raters were recruited as participants and randomly assigned to one of four conditions in a 2x2 factorial design. Factor A was a manipulation of the number of dimensions (7 vs. 2) to be considered simultaneously. Factor B was the additional requirement (or lack thereof) to rate the performance of standardized actors as well as the candidate being assessed. All participants were asked to observe videotapes of 3 unscripted clinical performances while verbally identifying relevant behaviours and rating the performance using the assigned scale.

Participants

To be eligible, participants must have held preceptor and/or faculty appointments in a school of health, with clinical experience and responsibilities for assessing clinical performance in simulation and/or workplace-based settings.

Materials

The three videos displayed unscripted clinical performances involving individual paramedic candidates attending to a patient (a mannequin) with a decreased level of consciousness secondary to a lethal arrhythmia who deteriorates to cardiac arrest. The candidate was expected to take a leadership role in assisting two lower trained first responders (standardized actors) with patient care. The actors were trained to appear distressed about the circumstances while varying in their willingness and ability to participate in patient care. Videos were selected from an existing pool of 80 available to the research team from a previous study (Tavares, et al. 2012). We randomly selected 1 from each of three performance groups (high, average, low) based on previously assigned ratings. Similar to OSCE encounters, each video presented a different candidate encountering the same clinical scenario with a set time limit (9 minutes).

To create the 7D condition, we used a 7-dimension (situation awareness, history gathering, patient assessment, decision making, resource utilization, communication, procedural skill) global rating scale that had previously undergone rigorous development and construct validation procedures (Tavares, et al. 2012, Tavares 2013). To create the 2D condition, we modified the same scale to include only two (history gathering and procedural skill) of the original seven dimensions. The 2 dimensions were selected based on their conceptual differences and having previously identified relatively low inter-item correlations (r=.66) (Tavares, et al. 2012). Other pairs of dimensions (e.g., History Gathering and Communication) also demonstrated evidence of empirical distinctiveness (i.e., inter-item correlation of .63) but conceptually seemed more related than the pairing we selected. Selecting dimensions that were relatively uncorrelated allowed us to prompt consideration of distinct aspects of performance. The scale includes definitions for each dimension and was scored using 7-point adjectival scales anchored with descriptive terms and definitions from 1=unsafe "…performance compromised patient care…unsuitable for supervised practice or progression" to 7=exceptional "…demonstrates high standard of performance…highly recommended for independent practice or progression".

In an effort to further manipulate rating demands, half of both the 7D and 2D groups were asked to assess the performance of the simulated colleagues (i.e., standardized actors). To do so they were provided with a modified version of the Maastricht Assessment of Simulated Patients (MaSP) rating scale (Wind, et al. 2004). The scale in its original form includes a total of twenty-one items with ten allocated to authenticity of the performance, ten allocated to feedback and one global overall performance score. We eliminated the feedback and global items and modified the scale to accommodate standardized healthcare professionals rather than patients. For example, where statements in the original scale were specific to portraying a clinical condition (e.g., "the SP might be a real patient"), revisions were made to reflect the context in this study (e.g., "the SA might be a real first responder"). While such alterations might change the psychometric properties of the MaSP, our intention here was simply to use the instrument to manipulate the rating demands placed on the raters (i.e., we were not interested in the psychometric defensibility of the rating tool).

According to Paas, mental load is defined by (a) task demands (i.e., difficulty independent of participant characteristics), (b) mental effort (the amount of mental resources allocated by the individual) and (c) perceived performance (Paas and Van Merriënboer 1994). Therefore, as a manipulation check, we created a self-report task measure with 3-items. Task demand and mental effort were scored with 9-point Likert scales anchored on the odd values from 1 (not demanding or no effort) to 9 (extremely demanding and requiring extreme effort). The

performance item referred to ability to identify all behaviours or events on the assigned GRS and was anchored using incremental proportions from less than 20% (of behaviours present in each video) to greater than 90% on a 9-point scale.

Procedure

Following collection of informed consent and randomization, participants were oriented to their assigned GRS, the modified MaSP (if relevant), details regarding the case, the task measure and a general overview of the process. Participants were instructed to verbally identify all behaviours or events that were associated with the dimensions included on their assigned GRS, to specify the relevant dimension for each observation and to indicate whether the behaviour was suggestive of competence or incompetence. Participants in the additional task group were instructed to place equal importance and consideration on all 3 observable individuals.

Prior to data collection, all participants were given an opportunity to practice using an unrelated video. A research assistant guided and continued the practice session until it was clear that the participant understood and could apply the instructions.

The sequence of study videos was randomized within group. Participants were not allowed to pause or rewind the video. Immediately following each video, participants were given a maximum of 5 minutes to complete the GRS and, if applicable, an additional 5 minutes to complete the modified MaSP. Measures of mental load and performance were taken at the three, six and eight minute marks of each video. A research assistant paused the video, allowed the participant to complete their thought if doing so would have been interruptive, and asked the participant to complete the task measure. Participants then returned immediately to the videos without rest or opportunity for discussion.

All participants were interviewed at the very end of the exercise using a one-on-one, face-to-face, semi-structured protocol with trained research assistants. The interview questions were designed to determine participants' strategies to resolve or manage instances in which they may have felt overwhelmed by the load associated with the task.

Participants' identification of dimension relevant behaviours during the rating tasks and interviews were captured using a wireless audio recorder followed by verbatim transcription in preparation for analysis. The GRS, mental load measures and MaSPs were paper-based.

Outcome Measures and Analysis

(a) Number of Dimension Relevant Behaviours Identified

As assessments are dependent on behaviours observed, we used the number of dimension relevant behaviours identified as an indicator of the sample of behaviours raters thought to be particularly useful in enabling their judgments for a particular dimension.

For the two dimensions common to all groups (history gathering and procedural skill), a criterion was created by having two experienced raters identify all the behaviours observable in each video that aligned with the corresponding dimensional definitions. They independently considered one dimension at a time, pausing and rewinding the videos as necessary until they were confident that all behaviours had been identified. They then met to discuss discrepancies

and reach consensus on a final criterion. Any behaviours not identified by the experts, but identified by the participants as being relevant, were reviewed by the two experts and searched for in the corresponding videos. Any behaviours not previously identified, but confirmed to be present in the videos and aligned with the definitions included on the GRS were added to the criterion and transcripts were then rescored using the revised criterion as necessary.

Using the proportion of dimension relevant behaviours identified (given our criteria above) for the two dimensions common to both groups, we conducted a 2 (7D vs. 2D) x 2 (Additional Task vs. No Additional Task) x 3 (Video) ANOVA. We recognized that participants, particularly in the 7D conditions, may have identified a sample of behaviours included on our criterion, but not assigned the behaviour to one of the 2 dimensions included in our analysis. Ignoring this data may have led to misleading conclusions. Therefore, we repeated the analysis after crediting participants with identifying the sample behaviour (even if assigned to another dimension) to test the robustness of our results.

We hypothesized that being asked to attend to seven dimensions and the instruction to complete additional tasks would decrease the number of dimension relevant behaviours identified. These analyses were conducted using SPSS Ver. 19.

(b) Reliability

Using the GRS scores for the two dimensions common to all groups, we used Generalizability Theory to calculate inter-rater reliability and internal consistency (Brennan 2001). Generalizability theory uses ANOVA to parcel variance in scores into that which is attributable, in this design, to the candidates, the raters, the items, and the various interactions between these variables. Because item contributed little variance, Classic Test Theory formulations were used to calculate 95% Confidence Intervals around the inter-rater reliability estimates. We hypothesized that being asked to attend to seven dimensions would result in a decrease in inter-rater reliability relative to being asked to attend to only two dimensions. An impact on internal consistency was uncertain as this metric is commonly high in these types of assessments and sub-optimal processing could theoretically increase the halo effect (erroneously high correlations between raters) were ceiling effects not limiting. Reliability analyses were conducted using G-String Ver. IV.

(c) Post Task Semi-Structured Interview

To explore the cognitive strategies raters engage when they are prompted to maintain a broad focus in generating their assessments, we analyzed interview data collected from 7D participants using an inductive thematic content analysis(Elo and Kyngäs 2008). Audio recordings were transcribed verbatim and checked for accuracy. Initially we used open and inductive line-by-line coding (provisional codes) (Charmaz 2006) using either direct statements made by participants or interpretation. These initial codes and some transcripts were reviewed by a second researcher to ensure they were derived appropriately from the data and revised as necessary. Two researchers (WT, SG) then compared the codes and grouped them together as appropriate. We returned to the data repeatedly to search for additional codes, recode segments as necessary, assess for fit of our codes and further refine the codes. We repeated the process until it was evident that the themes were adequately representative of (and clearly derived from) the data. QSR NVivo was used to store and manage our qualitative data.

Results

A total of 85 raters were recruited across Ontario and Nova Scotia Canada to participate in this study (21 or 22 per cell of the 2x2 design). Table 1 provides details regarding participant demographics.

Manipulation Check

To determine whether our manipulation had the intended effect, we calculated individual mean scores for each of the 3 items included on our working load measure (table 2). ANOVAs revealed consistent effects of number of dimensions assigned (7 vs. 2), but no main effect of the additional rating task and no interaction between these variables.

Number of Dimension Relevant Behaviours Identified

Our criterion for videos A, B and C included 40, 34 and 32 dimension relevant behaviours for "procedural skill" and 19, 10 and 11 behaviours, respectively, for "history gathering". The mean proportion of dimension relevant behaviours identified per group and by dimension are provided in table 3.

For the dimension "procedural skills", a 2 (dimension) x 2 (task) x 3 (video) ANOVA with dimension and task set as between subjects factors and video as set as a within subject factor, revealed a main effect of dimension (7D mean=36.2 (95%CI=32.5-40.0); 2D mean=22.4 (95%CI=20.8-26.3); $F(1,80)=94.3, p<.05$, but no effect of additional task (no additional task mean=32.0 (95%CI=27.8-36.1); additional task mean=27.8 (95%CI=24.2-30.1); ($F(1,80)=4.9, p>.05$) and no interaction ($F(1,80)=.32, p>.05$). Similarly, for the dimension "history gathering", the same analysis revealed a main effect of dimension (7D mean=38.6 (95%CI=33.5-42.9); 2D mean=24.0 (95%CI=21.1-26.9); $F(1,80)=31.6, p<.05$), but no effect of additional task (no additional task mean=32.2 (95%CI=27.7-36.1); additional task mean=30.4 (95%CI=25.2-34.6); ($F(1,80)=.43, p>.05$) and no interaction ($F(1,80)=.52, p>.05$).

The inconsistent or non-existent effect of the "additional tasks" variable both here and in the manipulation check led us to collapse data across that variable to maximize the sample size considered in subsequent analyses.

Reliability

The 2D group achieved higher inter-rater reliability (G=.11 for a single observation and .62 when averaged over 2 items and 10 raters) relative to the 7D group (G=.07 and .48, respectively). Internal consistency was similar at G=.39, and .40 in the 7D and 2D conditions, respectively (.71 and .78 when averaged over 2 items and 10 raters) . See table 4 for a summary of the variance components, the percentage of total variance attributable to each facet, and the formulae used to calculate each coefficient.

Semi-Structured Interviews

Based on the results presented above (i.e., manipulation check and consistent findings of a main effect of dimension) we used data from interviews gathered from participants who were asked to assess candidates using 7 dimensions to address our second research question. This resulted in a total of 43 transcripts. Our analysis of those transcripts revealed a number of themes that

coalesced into three main groupings; (a) sources of load, (b) factors that may promote load reduction, and (c) active point-in-time strategies participants applied in response to high demands.

Sources of Load

Sources of load could be grouped as related to intrinsic or extraneous load. Intrinsic load (i.e., mental effort directly related to the rating task itself) included (a) difficulty assigning behaviours to predefined dimensions, (b) difficulty with behaviours that may be included in more than one dimension; (c) poor candidate performances including number of errors and/or organization, (d) considering multiple behaviours or dimensions simultaneously, and (e) too many visual/auditory stimuli. Extraneous load (i.e., mental effort not related to the rating task) included (a) unrelated visual/auditory distractions and (b) the complexity associated with the interaction between relevant and irrelevant events or behaviours. See table 5 for a summary of the sources of load, a description for each and sample quotes.

Factors that may Promote Load Reduction

Factors that may contribute to load reductions (mainly intrinsic load) included (a) familiarity with case, (b) familiarity with the dimensions and (c) repetition. See table 6 for a summary of these factors, a description for each and sample quotes.

Active Strategies to Reduce Load

Participants reported many active point-in-time strategies they engaged to help them complete the rating task that seemed consistent with an attempt to reduce intrinsic and/or extraneous load. For instance, participants reported engaging in a process of *prioritization*. Rather than pay attention to all seven dimensions, they would spontaneously reduce demands by selecting behaviours or focusing on dimensions that they considered to be "most important". In other words, they would attempt (intentionally and unintentionally) to reduce the number of dimensions they considered in a manner consistent with the theory that led to our experimental manipulation. Notably, what was described as important varied among the participants. Another strategy to reduce intrinsic load was *simplification*. Here participants reduced the rating process to what was manageable by picking out what was "easiest" to identify, eliminating an entire spectrum of behaviours by "focusing only on negative behaviours", and/or resorting to only those behaviours that were "most obvious". Here too, the behaviours described as easiest or most obvious and whether or not raters considered only negative behaviours to the exclusion of others, was idiosyncratic. A third strategy involved the use of *cognitive aids*. Here participants created activities (e.g., cue cards, note-taking, referring often to the rating scale) to help manage the rating task. Whether this was an effective strategy or resulted in additional load is unclear.

To reduce what participants felt was extraneous load, they actively engaged in a process of *exclusion*. Raters who were prompted to perform the additional rating task and those who were not would intentionally "block out" the standardized actors. Importantly, however, some realized that they did so only after the rating task and upon reflection, indicating that simplifying strategies sometimes occurred without conscious awareness and despite clear instructions to consider elements that were excluded as an objective of the rating task. See table 7 for a summary of these categories, a description of each strategy and sample quotes.

Discussion

As rater judgment increasingly becomes recognized as valuable for assessment of clinical competence (Van der Vleuten, Schuwirth et al. 2010, Hodges 2013), the purpose of this study was to explore and understand the alignment (or lack thereof) between imposed rating demands and inherent human cognitive architecture. This parallel mixed methods study allowed us to experimentally explore the effect of manipulations in rating demands while also uncovering potential mechanisms and insights surrounding our findings and that of other related research(Tavares and Eva (submitted)). Consistent with our hypothesis and as predicted by working memory(Baddeley 1992), information processing(Wickens and Carswell 2006) and cognitive load theories(Sweller 1988), as the number of dimensions increased, the ability to identify dimension relevant behaviours declined, resulting in fewer and more variability of behaviours observed. That is, having to consider multiple dimensions or competencies simultaneously may increase mental load to the point of exceeding inherent cognitive resources and impairing performance on intended goals. As judgments regarding clinical competence are ultimately based on behaviours observed (detected / selected and/or processed), these results were consistent with or may have led to the observed decreased ability to consistently differentiate between performances.

When we explored the cognitive strategies raters engaged in under high demand conditions to better understand these results, we learned of numerous sources of load, factors that might lead to a reduction in load in the future and point-in-time active strategies raters engaged to manage or reduce load. As the majority of these active strategies appear to be idiosyncratic and designed to reduce effort rather than optimize performance, the results of this study support earlier research suggesting rating demands may contribute to rater error(Williams, Klamen et al. 2003, Melchers, Kleinmann et al. 2010, Tavares and Eva 2012) and begins to shed light on underlying cognitive behaviours that may explain some of the challenges observed in rater-based assessments. In particular, raters engaged a process of *prioritization* and/or *simplification.* They actively reduced intrinsic load (i.e., rating demands) by focusing on dimensions that they felt were most important when all could not be considered simultaneously. In a recent study exploring sources of variability in assessors' judgment, Yeates et al., observed similar behaviours in that assessors appeared to focus (and comment) on different aspects of performance despite viewing the same performances(Yeates, et al. 2012). The results of both studies suggest raters may ultimately form judgments "pertain[ing] to different (rather than the same) performances" when assessing the same candidate (Yeates, et al. 2012). It is not entirely clear why some dimensions were prioritized over others, but this act of prioritization has explanatory relevance. Simplification involved selecting the "easiest" and "most obvious" behaviours as opposed to the most relevant or the most comprehensive information. This demand avoidance or effort to align rating demands with cognitive resources may be a contributing source of variability or rater error resulting in differences in rater judgments as evidenced by our reliability results.

Finally, raters reported strategies to minimize what they perceived as extraneous load. For instance, when rating demands were high, raters reported "shut[ting] out" the standardized actors (despite instructions to the contrary for some) to allocate mental resources to the candidate. This may help explain the general lack of effect we observed of presenting an additional task. While such strategies may be effective in reducing cumulative load, applying similar strategies in non-

experimental settings, particularly in work-based assessment may be difficult or inappropriate. For instance, it would be difficult and sometimes dangerous to "shut out" details of a patient's presentation or environmental factors such as the risk of physical harm caused by the surrounding environment or the actions of team members. As a result, we cannot rule out the notion that extraneous demands of some forms might influence rater performance even though we did not observe consistent effects on this factor.

These cognitive strategies are analogous to and supported by instructional design research that draws on similar frameworks (Paas, et al. 2003; Paas, et al. 2010; Van Merriënboer and Sweller 2010), but has important differences. In instructional design research, strategies are designed to present information in a way that reduces load on learners (targeting both intrinsic and extrinsic load where appropriate). For example, "weeding" which refers to the elimination of interesting but extraneous information, can be effective (with an average effect size of .90) (Mayer and Moreno 2003). Other strategies, such as simple-to-complex sequencing for instance, have also been shown effective for learning (Van Merriënboer, et al. 2003, Plass, et al. 2010). Applying similar strategies to performance-based assessments in the real world will not always be possible for the reasons named above (i.e., where one has little control over the setting in which the ratings are collected). In fact, attempts to do so would likely be at odds with calls advocating for greater authenticity in the clinical challenges we present candidates (Hodges 2003). The requirement to ensure comprehensive construct representation in the development of rating tools (Downing 2003; Streiner and Norman 2008) may be addressable by deliberately limiting the focus of any given moment of assessment while aggregating across competencies over time.

It is interesting to note that broadening of focus with a fairly subtle manipulation (i.e., increasing the number of items included on a rating scale) appears to have been detrimental to rating quality despite raters' reported efforts to reduce load and even though the effort to raise load through an additional rating task was ineffective. We speculate that the number of dimensions to be considered is an intrinsic load whereas rating the standardized actor is an extrinsic load and that it is more difficult to effectively deal with intrinsic factors.

This study contributes to a growing body of rater cognition research that identifies idiosyncrasies in rater judgment and their implications on assessment perspectives in health professions education. We revealed rating demands as a possible mechanism for rater idiosyncrasies; others have explored the role of biases(Williams, et al. 2003), contextual and social influences (Govaerts, et al. 2007; Gingerich, et al. 2011), rater performance theories (Govaerts, van de Wiel et al. 2011), differential salience, criterion uncertainty and difficulties with information processing or transformations (Yeates, et al. 2012), the role of alignment between rating tools and the way raters inherent conceptions (i.e., schemas) regarding competence(Ginsburg, et al. 2010; Crossley and Jolly 2012) and motivational issues(Murphy, et al. 2004). While considering these studies one must keep in mind that idiosyncrasy does not necessarily mean inaccuracy. That is, the judgments raters generate may be different for a variety of reasons, but may be equally meaningful (Landy and Farr 1980) and may, as a result, become most powerful when combined (Eva and Hodges 2012; Hodges 2013). The results of this study suggest this is an especially valuable perspective when construct representation or contexts of assessment exceed inherent cognitive resources or preclude the reduction of rating demands on raters. From a psychometric perspective, efforts to minimize validity threats attributable to rater error may be

limited without further considering the inherent demands imposed by various rating tasks and the alignment (or lack thereof) with inherent rater cognitive capacity.

Limitations

There are several limitations to consider in this study. First, we used a simulation-based context involving only one case (a cardiac patient deteriorating to cardiac arrest). Therefore, generalizing our findings to other types of cases, particularly those with fewer distractions, may be limited. Secondly, we applied a global rating scale as opposed to a checklist. Using a checklist instead of a global rating scale might yield different results in either direction, as it may theoretically reduce or increase mental workload. It's simply unknown and unpredictable based on these findings. The decision to use a global rating scale in this study was based on the current direction of the assessment literature, which has increasingly supported the use of GRS and rater judgment over checklists in similar contexts (Norman, Vleuten et al. 1991, Vleuten 1996, Crossley, Humphris et al. 2002). The act of verbally stating dimension relevant behaviours, including specifying the dimension of relevance and whether they were suggestive of competence or incompetence, without doubt resulted in additional mental work for our participants and may have impacted the overall rate at which behaviours were identified. However, this manipulation was applied equally to all participants and, therefore, does not threaten internal validity. Finally, asking participants to reflect on their strategies may be inherently limited (Nisbett and Wilson 1977), but our goal was to inquire about participants' perceptions rather than what may have been the right or wrong thing to do.

Conclusions

Many of the theoretical foundations we used as our conceptual framework have largely been applied to learning and instructional designs where researchers and designers have the ability to manipulate or control load induced by the "stimuli" on the "end user". In contrast, many trends in assessment and the surrounding validity arguments advocate for things that place greater demands on raters including the importance of authenticity and comprehensive construct representation. Workplace-Based Assessment, for example, involves presenting candidates with clinical challenges that require the integration of multiple competencies in real world settings. Assuming authenticity is important, manipulating the stimuli in this setting is impossible and arguably of limited value, leaving the end user (i.e., the rater) needing to find ways to manage the load imposed. When presented with high rating demands (i.e., 7D vs. 2D), raters experience difficulty in identifying and considering dimension relevant behaviours (seemingly without awareness of the performance deficiency) and appear to engage in idiosyncratic effort or load-avoidance strategies that may satisfy rather than optimize rater performance. The results of this study suggest that without further considering the demands imposed by assessment design, raters may be overwhelmed and likely to engage in processes that can negatively affect the validity of the assessment practices.

Table 1: Participant demographics by group.

Item	7D with no Additional Tasks n=21	7D with Additional Tasks n=22	2D with no Additional Tasks n=21	2D with Additional Tasks n=21
Age, Years (Mean, SD)	36(7.8)	39(10)	39(9.1)	34(8.1)
Male (%, n)	14	13	14	14
Female (%, n)	7	8	7	7
Highest Education (%, n)				
College	48%(n=10)	68%(n=15)	52%(n=11)	76%(n=16)
University	48%(n=10)	32%(n=7)	38%(n=8)	9%(n=2)
Graduate School	0.5%(n=1)	0%(n=0)	10%(n=2)	14%(n=3)
Professional Designation				
Primary Care Paramedic	48% (n=10)	46%(n=10)	43%(n=9)	62%(n=13)
Advanced Care Paramedic	52% (n=11)	54%(n=12)	57%(n=12)	38%(n=8)
Professional Clinical Experience (Years; Mean, SD)	7.2(6.6)	10.1(7.4)	8.1(7.8)	6.2(6.2)
Familiarity with GRS (Mean, SD)	5.3(3.1)	3.0(3.0)	5.1(4.1)	4.2(3.4)
Previous Rater Training	3	0	1	3

Table 2: Summary of individual means (including 95% confidence intervals) and inferential statistics for self-reported task demands, mental effort and perceived performance associated with the rating task

	Main Effect: Dimensions		Main Effect: Additional Task	
	Means (95% CI)	Statistics	Means (95% CI)	Statistics
Demands	7D = 6.1 (5.7-6.5); 2D = 5.2 (4.8-5.6)	F(1,80) = 7.7, p = < .01	With Additional Task = 5.9 (5.5-6.3) No Additional Task = 5.5 (5.0-5.9)	F(1,80) = 1.7, p = .19
Effort	7D = 6.1 (5.7-6.6) 2D = 5.4 (5.0-5.8)	F(1,80) = 6.0, p = .02	With Additional Task = 5.9 (5.5-6.4) No Additional Task = 5.5 (5.1-6.0)	F(1,80) = 2.3, p = .13
Performance	7D = 5.1 (4.7-5.5) 2D = 5.7 (5.2-6.1)	F(1,80) = 3.7, p = .06	With Additional Task = 5.2 (4.8-5.6) No Additional Task = 5.6 (5.2-6.0)	F(1,80) = .1.9, p = .18

Table 3: Mean proportion (across all 3 videos) of dimension relevant behaviours identified by dimension.

	History Gathering Skill				Procedural		
	7D	2D			7D	2D	
No Additional Tasks	25.6	38.8	32.2	No Additional Tasks	24.7	39.2	32.0
With Additional Tasks	22.3	38.4	30.4	With Additional Tasks	22.3	33.2	27.8
	24.0	38.6			22.4	36.2	

7D = group assigned seven dimensions; 2D = group assigned two dimensions

Table 4: Inter-rater reliability and internal consistency results including individual variance components and percentage of total variance for each facet for 7 dimension and 2 dimension groups.

	7D		2D	
Effect	**Var.**	**% of Total Var.**	**Var.**	**% of Total Var.**
Video	0.09	6.6	0.17	11.1
Rater	0.24	17.8	0.30	19.6
Rater x Video	0.43	31.9	0.44	28.8
Items	0.00	0.0	0.00	0.0
Item x Video	0.00	0.0	0.00	0.0
Rater x Items	0.07	5.2	0.17	11.1
Video x Rater x Items	0.52	38.5	0.45	29.4
Inter-Rater Reliability	G = .07; G_n = .48		G = .11; G_n = .62	
Internal Consistency	G = .39; G_n = .71		G = .40; G_n = .78	

7D = groups assigned seven dimensions; 2D = groups assigned two dimensions; Var. = Variance component.

G Formulae:

Inter-rater Reliability: [Var(v) + Var(vi)] / [Var(v) + Var(r)+Var(i)+Var(vr) + Var(vi) + Var(ri) + Var(vri)]

Internal Consistency: [Var(v) + Var(vr)] / [Var(v) + Var(r)+Var(i)+Var(vr) + Var(vi) + Var(ri) + Var(vri)]

G_n Formulae:

Inter-rater Reliability: [Var(v) + Var(vi)/2] / [Var(v) + Var(r)/10+Var(i)/2+Var(vr)/10 + Var(vi)/2 + Var(ri)/20 + Var(vri)/20]

Internal Consistency: [Var(v) + Var(vr)/10] / [Var(v) + Var(r)/10+Var(i)/2+Var(vr)/10 + Var(vi)/2 + Var(ri)/20 + Var(vri)/20]

Table 5: Results of semi-structured interviews describing sources of load for participants in high load rating conditions with descriptions for each, example quotes and supporting theoretical foundations.

Sources of Load	Description	Example Quotes
Intrinsic sources		
Assigning of Behaviours to Predefined Dimensions	Participants reported difficulty associated with aligning their observation of a behaviour with predetermined / predefined rating scale dimensions. This occurred when trying to align one behaviour with one dimension, but also with behaviours that appeared to align with more than one dimension.	#160: *"the difficulty was just finding where they [behaviours] fit, what slot [category / dimension] they fit into. Whether it was patient assessment or communication or whatever, that was the hardest part…"* #170: *"I felt it was difficult to divide each skill that the trainee was doing into a separate category [dimension] because some of them can fall under uh many of them."*
Candidate Performance Characteristics	The pattern of behaviours exhibited by candidates appeared to generate mental work. Participants reported those candidates with more errors and less organization for instance (based on the rater's standards) for instance, as more difficult.	#136: *"When the trainee was more organized I found it was easier… because it just goes with the flow of what you would want to do."* #137: *"and I definitely found differences in you know video B for me. Her performance was so much easier to evaluate because there seemed less wrong in it."*
Multiplicities of Focus	Participants reported difficulty with having to divide attention across multiple behaviours and/or dimensions.	#148: *"And then you're also thinking about the two [standardized] actors… It wasn't just focused on the candidate and going over what the candidate is doing. You also had to think about what the other two [standardized] actors are doing. So it kind of made it hard to just focus on the candidate and assess them"* #115: *"It was a little bit difficult because. It's really hard to pay attention to everything that's going on…the things that make it difficult in my mind is that I don't know what I'm not seeing. So I know I'm missing something because you pay attention to many little things, and you can't focus on everything. I, I, I can't look at everything at the same time."*

Too many auditory/visual stimuli	Refers to perceptual limitations given the amount of auditory and visual information raters were presented with (in contrast to the division of attention)	#101: *"Too many inputs, too many things to keep track of at once…mainly because there was so many different things to watch using different skills…sensory overload"* #126: *"I found just uh trying to catch everything and looking at everything that's going on was really hard, trying to focus on all the different aspect you had to look at, you had to listen for did they do the right, uh did they ask the right questions, are the delegating properly, are they cluing in to things, are they noticing things in a timely manner are they…are their assessments appropriate, are they accurate. It was really hard, you know especially with the noise and all that, I know I missed some stuff"*
Extraneous sources		
Unrelated auditory/visual distractions.	Participants reported the pull of cognitive resources toward unrelated (i.e., extraneous) content.	#107: *"Once the, specifically the auditory distractions were removed, then it was very, it was a lot easier cause I could focus on what the candidate was doing. On a skill-by-skill basis because you know I could just watch what they're doing. But when there's too much distraction in the background I found I would be distracted by what they [standardized actors] were saying, and it would distract me from watching what the candidate was doing, even though it was all right in front of me it's still, it keeps shifting in focus, you know."*
Complexity associated with interactions between relevant and irrelevant data	The act of simply eliminating what raters believed was extraneous at times was limited because of the interaction between what they identified as meaningful and meaningless.	#157: "*…also I'm getting distracted cause now I'm listening to her [the candidate] and trying to figure out what he [a standardized actor] can and can't do."* #121: *"…It is difficult as an assessor to um to continually try and tune that [standardized actors] out and yet not ignore the fact that he [the candidate] should have been dealing with them as well…it was difficult, but I listened very carefully to what the candidate was asking, the response that he was getting whether they were following through with his instructions or whether he [the candidate] was um listening to what was being told to him….Um and then what he did with that info."*

Table 6: Results of semi-structured interviews describing factors that may promote load reduction for participants in high load rating conditions with descriptions for each, example quotes and supporting theoretical foundations.

Factors that may Promote Load Reduction	Description	Example Quote
Familiarity with Case	Refers to the reduction in mental workload associated with greater as familiarity with the case. It appears raters build schemas related to the case, which may reduce workload.	#141: *"I think the, you know the first one because you are really kind of engulfed in everything and you are trying to pay attention to the whole situation, you uh you lose focus on, uh maybe the evaluating part right, because you are paying attention to everything that's going on. Whereas by the, by video C for me, I felt you know I didn't have to necessarily pay as much attention to the bystanders, as much attention to the patient, where they're going"* #144: *"With each video I feel like I I've kind of was able to move around the screen or in my head move around from point to point a little bit faster, a little bit faster so certainly the more repetitive it got.. I I felt a little bit more comfortable managing these different categories and aspects. But it probably was just a little bit more familiar, familiarity with the scenario so I knew what was coming, so I was prepared to look for certain things."*
Familiarity with Dimensions	Refers to the reduction in mental workload associated with greater as familiarity with the dimensions included on the GRS. It appears raters build stronger schemas related to the dimensions, which may reduce workload.	#107: "*As I got more comfortable with the different categories I thought it was easier. The first time I think you tend to choose three or four categories that you will just focus on and to the detriment of the other three. But as you become more comfortable with those first three or four that you've picked then you start pulling the other ones in to the scheme of things.*"
Repetition	As participants were able to experiment with the interaction between cases and dimensions and observe various candidate	#106: *"As I went a long it got a little bit easier because I was able to adapt my standard approach of assessment to this new system. The first one was very challenging but the subsequent ones became easier and easier."*

	performances, they seemingly built schemas related to this interaction, which appeared to reduce workload.	#177: *"It seemed to get easier, kind of, through each video. It is just knowing, kind of, what you were initially looking for, able to pick up on things um (on the) last one versus the first one…it didn't seem to be quite as mentally demanding as you went through as it was the first time."*

Table 7: Results of semi-structured interviews describing actives strategies for participants in high load rating conditions with descriptions for each, example quotes and supporting theoretical foundations.

Active Strategies to Reduce Load	Description	Example Quotes
Prioritization	Refers to a cognitive strategy aimed at reducing workload by idiosyncratically selecting fewer dimensions (or individuals observable in the video) to observe and rate.	#155: *"This was a lot of categories [dimensions] for me. I found that I just stuck to a few certain categories. That is what made it easier for me."* #132: *"I find I was obviously focusing more on... like I found myself (focusing on) the procedural one, communication, recourse utilization um and sometimes decision… I found that during the scenario I was utilizing more (of one) than the other."* #107: *"The ones I chose were um situational awareness because the situation was something that really uh it was driving the call at times…procedural skills was an important one for me…uh resource utilization was another one…and communication. I think the one I used the least was decision making"*
Simplification	Refers to the active cognitive strategy of making work (i.e., the rating task) easier.	#106: *"And my standard approach is to pick out the critical and the more um obvious errors verses like the um smaller minute errors. I had to make it a constant adjustment to, to get down to like being a little bit more nitpicky with some of the minor details versus looking at the major details."* #115: *"A couple of times I did feel overwhelmed and what would usually happen then is I would just kind of shut down a little bit and look at, look for things that were obvious, instead of trying to dig, dig, dig"*
Cognitive Aids	Refers to the active strategy of searching for and/or using what participants believed would aid them in managing the workload (presumably when internal strategies	#134: "*But it's hard because, maybe we might need more guidance, or maybe a sort of a cue card stating this is exactly what you're looking for etcetera….I re-visited the headings myself that I had written down over and over again. And actually read some of the words within the headings [GRS definitions] as well, just to keep referring myself, and keep giving*

	were limited or ineffective).	*myself constant feedback as to what my task was. "* #137: *"I made short form notes that I'm more accustomed to reading to reference quickly during the scenario…I made point form notes of that so that when I needed to give feedback all I had to do was glance down and I could try to remember what category that would fall under."*
Exclusion	Refers to the idiosyncratic elimination of certain stimuli available in the rating task. At times this was inappropriate and at other times unintentional.	#177: "*Um trying to focus more on the actions of the candidate and blocking out the background uh you know treating the supporters [standardized actors] kind of like you would my own teenagers. Blocking out the background, the unnecessary and trying to get away from hearing that kind of information. Focusing on you know task at hand and how they actual assessed and treated the patient*"

References:

Baddeley, A. (1992). Working memory. *Science.* 31. Vol. 255 no. 5044: 556-559.

Brennan R.L. (2001). Generalizability theory. New York, NY. Springer Verlag.

Byrne A, Tweed N, & Halligan C. (2014). A Pilot Study of the Mental Workload of OSCE Examiners. *Medical Education*, *48*(3), 262-267.

Charmaz, K. (2006) Constructing grounded theory: A practical guide through qualitative analysis. London. Sage Publications Ltd.

Creswell, J.W., & Clark, V.L.P. (2011). Designing and conducting mixed methods research. Thousand Oaks. CA. Sage Publications Ltd.

Crossley, J., Humphris, G., & Jolly, B. (2002). Assessing health professionals. *Medical Education*. *36*(9), 800-804.

Crossley J., & Jolly B. (2012). Making sense of work-based assessment: ask the right questions, in the right way, about the right things, of the right people. *Medical Education*. *46*(1), 28-37.

DeNisi A. (1996). A cognitive approach to performance appraisal: A program of research: Routledge.

Downing, S., & Yudkowsky, R. (2009). Assessment in Health Professions Education. New York NY. Taylor & Francis.

Downing, S. (2005). Threats to the validity of clinical teaching assessments: What about rater error? *Medical Education*. *39*(4), 353-355.

Downing, S. (2003). Validity: on the meaningful interpretation of assessment data. *Medical Education*. *37*(9), 830-837.

Elo, S., & Kyngäs, H. (2008). The qualitative content analysis process. *Journal of Advanced Nursing. 62*(1), 107-115.

Eva, K.W., & Hodges, B.D. (2012). Scylla or Charybdis? Can we navigate between objectification and judgement in assessment? *Medical Education*. *46*(9), 914-919.

Gingerich, A., Regehr, & G., Eva, K.W. (2011). Rater-Based Assessments as Social Judgments: Rethinking the Etiology of Rater Errors. *Academic Medicine*. *86*(10), S1-S7.

Ginsburg, S., McIlroy, J., Oulanova, O., Eva, K.W, & Regehr, G. (2010). Toward Authentic Clinical Evaluation: Pitfalls in the Pursuit of Competency. *Academic Medicine. 85*(5), 780-786.

Govaerts, M., van der Vleuten, C., Schuwirth, L., & Muijtjens, A. (2007). Broadening perspectives on clinical performance assessment: rethinking the nature of in-training assessment. *Advances in Health Sciences Education*. *12*(2), 239-260.

Govaerts, M., Schuwirth, L., van der Vleuten, C., & Muijtjens, A. (2012). Workplace-based assessment: effects of rater expertise. *Advances in Health Sciences Education*.*16*(2), 151-165.

Govaerts, M., van de Wiel, M., Schuwirth, L., & van der Vleuten, C. (2011). Workplace-based assessment: raters' performance theories and constructs. *Advances in Health Sciences Education. 18*(3), 375-396.

Herbers, J., Noel, G., Cooper, G., Harvey, J., Pangaro, L., & Weaver, M. (1989). How accurate are faculty evaluations of clinical competence? *Journal of General Internal Medicine*. *4*(3), 202-208.

Hodges, B. (2013). Assessment in the post-psychometric era: Learning to love the subjective and collective. *Medical Teacher, 35*(7), 564-568.

Hodges, B. (2003). Validity and the OSCE. *Medical Teacher*. *25*(3), 250-254.

Holmboe, E.S. (2004). Faculty and the observation of trainees' clinical skills: problems and opportunities. *Academic Medicine. 79*(1), 16-22.

Huwendiek, S., Mennin, S., Dern, P., Friedman Ben-David, M., Van Der Fleuten, C., Tonshoff, & B., Nikendei, C. (2010). Expertise, needs and challanges of medical educators: Results of an international web survey. *Medical Teacher. 32*(11), 912-918.

Kogan, J.R., Conforti, L., Bernabeo, E., Iobst, W., & Holmboe, E. (2011). Opening the black box of clinical skills assessment via observation: a conceptual model. *Medical Education*. *45*(10), 1048-1060.

LaMantia, J., Rennie, W., Risucci, D., Cydulka, R., Spillane, L., & Graff, L. (1999). Interobserver Variability among Faculty in Evaluations of Residents Clinical Skills. *Academic Emergency Medicine*. *6*(1), 38-44.

Landy, F.J, Farr, J.L. (1980). Performance rating. *Psychological Bulletin*. *87*(1), 72-107.

Mayer, R.E., Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist*. *38*(1), 43-52.

Melchers, K.G., Kleinmann, M., & Prinz, M.A. (2010). Do Assessors Have Too Much on their Plates? The Effects of Simultaneously Rating Multiple Assessment Center Candidates on Rating Quality. *International Journal of Selection and Assessment*. *18*(3), 329-341.

Murphy, K., Cleveland, J., Skattebo, A., Kinney, T. (2004). Raters who pursue different goals give different ratings. *Journal of Applied Psychology*. *89*(1), 158-164.

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., van der Vleuten, C., & Graaff, E. (1991). Pitfalls in the pursuit of objectivity: issues of validity, efficiency and acceptability. *Medical Education*. *25*(2), 119-126.

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist*. *38*(1), 1-4.

Paas, F., van Gog, T., & Sweller, J. (2010). Cognitive load theory: New conceptualizations, specifications, and integrated research perspectives. *Educational Psychology Review*. *22*(2), 115-21.

Paas, F., & Van Merriënboer, J. (1994). Measurement of cognitive load in instructional research. *Perceptual and Motor Skills*. 79(1), 419-430.

Plass, J., Moreno, R., & Brünken, R. (2010). Cognitive load theory: New York. NY. Cambridge Univ Press.

Streiner, D., & Norman, G. (2008). Health measurement scales: a practical guide to their development and use. Fouth ed. New York: Oxford University Press.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive science*. *12*(2), 257-285.

Tavares, W., Boet, S., Theriault, R., Mallett,e T., & Eva, K.W. (2012). Global Rating Scale for the Assessment of Paramedic Clinical Competence. *Prehospital Emergency Care*. *17*(1):57-67.

Tavares, W., & Eva, K.W. (2012). Exploring the impact of mental workload on rater-based assessments. *Advances in Health Sciences Education*. *18*(2):291-303.

Tavares, W., & Eva, K.W. (submitted). The Impact of Rating Demands on Rater Based Assessment of Clinical Competence.

Tavares, W., LeBlanc, V.R., Mausz, J., Sun, V., & Eva, K.W. (2013). Simulation Based Assessment of Paramedics and Performance in Real Clinical Contexts. *Prehospital Emergency Care. 18*(1), 116-122.

van der Vleuten C, Schuwirth L, Scheele, F., Driessen, E., & Hodges, B. (2010). The assessment of professional competence: building blocks for theory development. *Best Practice & Research Clinical Obstetrics & Gynaecology*. *24*(6), 703-719.

Van der Vleuten, C. (1996). The assessment of professional competence: developments, research and practical implications. *Advances in Health Sciences Education*. *1*(1):41-67.

Van Merriënboer, J.J., Kirschner, P.A., & Kester, L. (2003). Taking the load off a learner's mind: Instructional design for complex learning. *Educational psychologist*. *38*(1), 5-13.

Van Merriënboer, J.J., & Sweller, J. Cognitive load theory in health professional education: design principles and strategies. *Medical Education*. *44*(1), 85-93.

Wickens, C.D., Carswell, C. (2006). Handbook of human factors and ergonomics, Third Edition. Salvendy G, editor. Hoboken, NJ: Wiley.

Wickens, C.D. (2002). Multiple resources and performance prediction. *Theoretical issues in ergonomics science*. *3*(2), 159-177.

Williams, R., Klamen, D., & McGaghie, W. (2003). Cognitive, Social and Environmental Sources of Bias in Clinical Performance Ratings. *Teaching and Learning in Medicine*. *15*(4), 270-292.

Wind, L.A., Van Dalen, J., Muijtjens, A., & Rethans, J.J. Assessing simulated patients in an educational setting: the MaSP (Maastricht Assessment of Simulated Patients). *Medical Education*. *38*(1), 39-44.

Yeates, P., O'Neill, P., Mann, K., Eva, K.W. (2012). Seeing the same thing differently. *Advances in Health Sciences Education*. *18*(3), 325-341.

Chapter 6 - Passive vs. Immersed: Rater Performance under Different Load Conditions and the Impact on Feedback

Abstract

Introduction:

The validity associated with rater judgment in assessments of clinical competence is dependent on raters' cognitive performance. Rating demands have been associated with inter-rater reliability impairments but the influence of such demands on the utility of assessment for learning has yet to be examined. The purpose of this study was to experimentally explore the role of rating demands on feedback delivery during the assessment of clinical competence in simulation based settings.

Methods

Participants were randomly assigned to the role of "passive rater" who simply observed performance or "immersed rater" who played an active role in case simulation in an OSCE while assessing clinical performance using a 7-dimension global rating scale. The immersed condition was intended to increase extraneous demands associated with the rating task by having raters act as a standardized family member in the case. Immediately following, all participants were required to rate student performance and document formative feedback.

Results

A total of 20 raters (10 per group) assessed 20 candidates in a 10 station OSCE. Repeated measures 2 (rater type) x 10 (station) ANOVAs revealed that participants in both groups rated student performance equally, but those enrolled in the passive group gave more feedback (mean=15.6 statements per candidate (95% CI =14.7-16.6)) relative to immersed raters (mean=12.2 (95% CI=11.4-13.1); p=<.01). Decision-studies revealed inter-station reliability of G=.61 and .49 for the passive and immersed conditions, respectively, when the scores of 10 stations were averaged. Internal consistency in both conditions was equivalent at G=.84.

Conclusions

The results of this study contribute to a growing body of research examining rater cognition in the assessment of clinical competence by indicating that the requirement to take on an active role in a simulation-based assessment can decrease the amount of formative feedback that raters provide candidates even when their ratings indicate comparable degrees of concern about candidate performance.

Introduction

Drawing accurate conclusions regarding a practitioner's general level of competence is a complex and challenging process. The outcomes of decisions regarding competence have important implications for patient safety and public trust, but also learner development and candidates' career advancement. The broadening of competencies expected of health professionals, the way in which these competencies interact and the role of context are just some factors contributing to the difficultly. Frameworks commonly applied to structure and optimize assessments of competence (see Miller 1990 and Kane 1992, 2006) have emphasized the value of direct observation of candidate behaviours in simulation and/or work place based settings (Miller, 1990; Kane, 1992; Kane, 2006; Downing and Yudkowsky, 2009; Kogan and Holmboe, 2013). The authenticity associated with observing what candidates *do* in response to actual clinical challenges requires less of an extrapolation when inferring clinical competence or making predictions regarding future clinical performance. Deriving competence from clinical performances in this way is vitally dependent on raters' judgment.

Assessments of clinical performance can be an issue when they are reliant on rater inference or judgment because numerous problems have been identified. For example raters may be subject to observational inaccuracy, prone to numerous rater biases and may have difficulty identifying performance deficiencies (Herbers, et al. 1989, Noel, et al. 1992, Morgeson and Campion 1997, LaMantia, et al. 1999). As such, raters cannot be viewed as objective measurement devices simply transferring candidate behaviours to rating tools (Landy and Farr 1980). Recent work has shed insight into this process by adopting a cognitive capacity model wherein raters are considered to have a finite capacity for holding task-relevant information in working memory (Baddeley, 1992; Tavares and Eva 2012). The general argument is that many of the cognitive structures needed to complete assessments of competence are capacity limited and, if overloaded, can lead to impairments in rater performance (Tavares and Eva 2012). Byrne and colleagues have reported that the cognitive load placed on raters in fairly standard OSCE situations can be greater than what trainees experience during their clinical activities (e.g., during the induction phase of anesthesia delivery) (Byrne, et al. 2014). Two additional studies in health professions education that directly tested hypotheses derived from this model suggest that when raters' cognitive load increases by tasking them with considering multiple dimensions of performance simultaneously, indicators of rating performance including inter-rater reliability, the ability to discriminate between levels of performance, and the ability to identify dimension relevant behaviours, become impaired. (Tavares and Eva (submitted))

Interestingly, in both studies, *intrinsic* cognitive load manipulations (i.e., those that are an embedded part of the rating task, such as requiring 7 dimensions to be assessed rather than 2) showed a consistent effect of decreased inter-rater reliability whereas two *extraneous* load manipulations (i.e., additional, but relevant activities such as generating patient care plans or simultaneously rating the performance of standardized actors) showed no influence. Interview data suggested that raters made deliberate and seemingly effective attempts to minimize cumulative load by minimizing attention towards what they believed to be extraneous. Whether extraneous demands can routinely be ignored in this way or not has important implications because many performance-based assessments require raters to conduct additional extraneous tasks. For example, in controlled settings, the main difference between OSCEs and Simulated

Office Orals is whether the rater observes a performance or takes on the role of the patient when conducting the assessment. In more real world settings such as work-place based assessments raters are often required to divide their attention by interacting with other individuals during assessments, monitoring patient safety, or monitoring the performance of other team members. Cognitive load theory suggests that both intrinsic and extraneous load will influence a capacity limited system, but in studies we have conducted, we have yet to see an influence of extraneous load on the performance of raters. Nor have we identified studies elsewhere that have addressed this issue directly.

Studies exploring rating demands on rater-based assessments have primarily focused on the utility of the scores assigned, whereas the implications for learners have not been explored. Given the importance of feedback as a mechanism to improve performance, it would be problematic if increasing degrees of cognitive load decreased the guidance provided to candidates. Therefore, the purpose of this study was to explore the role of extraneous rating demands on the feedback provided during the assessment of clinical competence in simulation based settings by holding intrinsic demands constant and experimentally introducing extraneous demands.

Methods

Study Overview

This experimental study involved randomly assigning raters to either the role of "passive observer" or "immersed actor" in a simulation based, performance based exam. Participants assigned to the passive condition were tasked simply with observing and rating candidate performances using a 7 dimension global rating scale (GRS). Participants assigned to the immersed condition were similarly tasked with observing and rating candidate performances but were active participants in the case as they were required to play a standardized role and to guide the direction of the simulation as necessary. Immediately after rating student performance, all participants were asked to document the feedback they would deliver to support candidate development. Outcome measures included the amount of feedback given both in total and specific type. Research ethics approval was provided by Centennial College in Toronto, Ontario, Canada.

Setting

Data were collected in the context of an existing high stakes 10 station, 20 candidate, simulation based assessment of paramedic clinical competence (at the end of their training), using an OSCE format. Each station lasted 12 minutes and candidates were given 5 minutes to rest between stations. Each station involved a distinct clinical case that required candidates to demonstrate clinical skills such as history gathering, physical exam, interpreting assessment results, formulating working diagnoses and implementing emergency treatment plans. Cases were developed by Centennial College in Toronto, Ontario Canada and represented a wide variety of case types based on real clinical cases seen during paramedic practice. Each station included a standardized patient (or mannequin), a complete set of standardized equipment that was not specific to the case (or necessarily required to perform in that case), as well as contextually relevant props as needed. Because paramedics work in teams of 2, each station included a

standardized "partner" who was assigned to the case and assisted candidates in the performance without initiating actions themselves.

Participants

We used purposeful sampling to recruit raters for this study. Eligibility criteria were being an active clinician or educator in the field of paramedicine and having experience rating paramedic candidate performance in simulation or work based settings. No rater training was included as part of this study, either for the assessment of competence or provision of feedback to avoid confounding the outcome measurement. All raters received financial compensation for their participation.

Rating Tool

All participants used a 7-dimension GRS that included situation awareness, history gathering, patient assessment, decision-making, resource utilization, communication, and procedural skill(Tavares, Boet et al. 2012). This GRS was previously developed specifically for the assessment of paramedic clinical competence and has demonstrated evidence of both reliability and construct validity in similar settings(Tavares, Boet et al. 2012, Tavares 2013). It is the scale that was used in the high intrinsic load condition of previous studies. (Tavares, Ginsburg et al. 2014, Tavares and Eva (submitted)) For each of the dimensions included on the GRS, a definition is provided to promote a shared mental model across examiners. A 7-point adjectival scale accompanies each domain with each of the 7 points defined making reference to patient safety, performance standards and/or readiness for independent practice or progression. Finally, the GRS includes space for comments / feedback under each dimension.

Procedure

Following recruitment and informed consent, raters were randomly assigned to a specific station and to the role of passive observer or immersed actor, such that each station had one passive and one immersed rater who remained together for the duration of the assessment. The paired participants were provided with the same detailed description of the case, which included a case stem (i.e., call information for the candidate), history elements, patient characteristics (e.g., gender, age, and physical exam findings), details regarding how the case was to unfold and performance expectations.

Raters assigned the role of passive observer were required to complete the rating task while not being explicitly involved in the simulation in any way. They were present in the room where the simulation was taking place, along with the immersed rater, and were free to position themselves as needed to observe the case, but had no additional demands or responsibilities beyond observing and rating candidate behaviour. The candidates, the standardized patients and the individuals who played the role of standardized partner paramedics were instructed to avoid engaging the passive raters in any way.

In contrast, the immersed raters were tasked with the same rating requirements as the passive rater, but were additionally required to serve the role of standardized bystander (i.e., family member) and provide whatever information was necessary to the candidate during the clinical interaction. For example, it was the immersed rater's responsibility to relay patient details or history elements as a family member might be expected to do when asked by the candidate.

They were also expected to provide necessary data in instances when the limitations of simulation presented themselves (e.g., confirming the presence or absence of physical exam findings). Hence, the immersed rater moved "in and out" of standardized bystander/actor and simulation facilitator roles while simultaneously being responsible for rating clinical performance. These additional tasks were expected to increase extraneous mental workload for the immersed rater.

Regardless of role or station, participants were instructed to observe candidate performance and then complete the GRS without consulting with one another. Both groups of participants were also instructed to document dimension specific feedback that would result in maximal development opportunities for the candidate. To ensure participants were offered the same opportunity to assign scores and provide feedback, both groups were instructed to do so only after the case was completed. Because the immersed raters were not able to review the rating tool throughout the performance, passive raters were instructed to avoid using the scale during the performance. However, both groups were permitted to review the rating tool in advance of each performance. Five minutes was allocated between candidates for scoring and formative feedback. Even upon completion of their task, all participants were instructed to avoid sharing their scores or feedback with one another to prevent rater calibration over time.

Data Collection

Prior to beginning the assessment process, all participants completed a demographic form and rated their level of familiarity with the assigned global rating scale from 1 (no familiarity at all) to 10 (extremely familiar). All data were collected using paper-based versions of the GRS described above. Numeric data were transferred to an Excel spreadsheet and narrative comments were transcribed without identifying information in preparation for analysis.

Data Analysis – Leniency

To evaluate whether scores assigned by each group differed in some meaningful way (e.g., leniency or stringency), we compared scores for each dimension by rater type using one-way ANOVAs. For all statistical tests we used an alpha of .05 to determine significance.

Data Analysis - Feedback

The main outcome of interest was the amount and type of feedback given to candidates. To evaluate the effect of increasing extraneous mental workload demands on feedback, we analyzed textual data using deductive content analysis (Elo and Kyngäs 2008). Literature describing principles of effective feedback and associated theoretical models were used to derive a coding structure (Hattie and Timperley 2007, Shute 2008, Archer 2009). Statements were assigned to one of 6 categories: (1) feedback regarding a specific behaviour or task; (2) general feedback regarding a dimension of performance; (3) general statements regarding the individual candidate; (4) general statements involving the context; (5) statements containing specific recommendations for improvement; and (6) statements encouraging reflection. See table 1 for feedback type definitions. The coding structure was piloted using data that were available to the research team from previous performance based assessments and modified until the structure was determined to be comprehensive and stable.

Using de-identified data, coders first segmented the feedback on the basis of semantic features such that each segment represented a single thought, idea or statement. Each rating form with feedback documented, was then coded by 2 independent coders who met to discuss discrepancies. Where disagreement remained, an additional coder (WT) reviewed the arguments and assigned a final code. One of the coders (WT) also randomly reviewed portions of the data set to cross check codes and ensure adequate segmentation and application of codes. Given that this study was embedded in an existing assessment process and clinical performances were not recorded we were unable to determine the adequacy of feedback, but were able to measure and compare the amount of feedback provided by type using repeated measures ANOVA. For all statistical tests we used an alpha of .05 to determine significance.

Data Analysis – Reliability

We used scores assigned on the GRS to calculate inter-station reliability and internal consistency coefficients for each group (i.e., passive and immersed) separately. Given the robust nature of context specificity it was not anticipated that reducing mental workload would necessarily improve these reliability estimates because differences across station can result from differences in actual performance rather than rater limitations. In addition, the inter-rater reliability examining the agreement between raters in the passive and immersed groups was examined for the sample as a whole. All reliability analyses were calculated using generalizability theory and G-String Ver. 4 (Brennan 2001, Bloch and Norman 2012).

Results

A total of 20 raters (10 passive and 10 immersed) were involved in evaluating 20 paramedic candidates over 1 day. Overall, the passive and immersed groups were similar in gender distribution, years of clinical experience, familiarity with the GRS and previous rater training. See table 2 for demographic details for both groups.

Leniency

Using scores assigned by raters in each group by dimension, ANOVA revealed no significant differences in mean scores for 6 of the 7 dimensions (see table 3).

Feedback

The feedback entered on 400 GRSs resulted in a total of 5,580 statements coded across both groups. Across all 6 categories, passive raters provided 3,132 individual feedback statements and immersed raters provided 2,448. After visually reviewing the number of feedback comments provided by individual raters for outliers and finding none, a repeated measures 2 (rater-type) x 10 (station) ANOVA performed on the total number of statements revealed a significant main effect of rater type that indicated passive raters provided more feedback overall when compared to immersed raters; passive rater mean=15.6 (95% CI = 14.7-16.6) vs. immersed rater mean=12.2 (95% CI = 11.4-13.1); ($F(1,119)=47.9$, $p<.01$, $d=.66$). When we conducted the same analysis on each type of feedback independently, we found a main effect of rater-type for 4 of the 6 types of feedback (feedback regarding a specific behaviour or task, general feedback regarding the dimension, general statements regarding the individual and statements involving specific direction or recommendations). See table 4 for individual means, standard errors, 95% confidence intervals and p-values by feedback category.

Reliability Analysis

Using GRS scores assigned by participants in each group we used generalizability theory (Brennan 2001) to calculate variance components for the facets "candidate", "station" (which is confounded with rater in this design), "item" and their interactions. Generalizability coefficients (i.e., reliability) provide information regarding the degree to which to the scores assigned consistently differentiate between candidates as well as an indication of the degree of measurement error contributed by various sources. A higher proportion of variance attributable to candidate suggests better differentiation. See table 5 for details regarding individual variance components, the percentage of total variance attributable to each facet, d-study results and the formulae used to calculate each coefficient. For passive and immersed raters the internal consistency for a single item (i.e., the average correlation between individual items) reached G=.08 and .44 respectively. When scores were averaged over 7 items, the internal consistency in both conditions reached G=.84. The inter-station reliability of a single station was expectedly low G=.08 and .05 in the passive and immersed conditions, respectively, a number that increased to G=.61 and 0.49 when the scores were averaged across all 10 stations. When we combined the data from both groups to allow for the calculation of inter-rater reliability between passive and immersed raters, the inter-rater reliability was found to be G=.45 suggesting the raters in each condition generally scored the candidates in different ways despite evaluating the same performance.

Discussion

Modern assessment frameworks have emphasized the value of direct observation when drawing conclusions regarding clinical competence (Govaerts and Vleuten, 2013; Kogan and Holmboe 2013). This has led researchers to explore how raters form judgments in these settings to understand their challenges and limitations (Govaerts, et al. 2011, Tavares and Eva 2012, Yeates, et al. 2012). Recent work applying a cognitive capacity model has suggested that excessive rating demands, in particular high intrinsic demands, can impair rating performance in simulation-based settings (Tavares and Eva 2014). The same has not been true with extraneous demands, in part because raters may have employed avoidance strategies to minimize such demands. In realistic settings, where patient safety concerns, for example, are real and draw on cognitive resources, applying similar strategies in practice may not be feasible or appropriate. The purpose of this study was in part to further explore the role of extraneous rating demands on rater performance. We held intrinsic demands constant and experimentally manipulated extraneous load in a manner that could not easily be ignored. The second purpose was to explore the impact of rating demands on the provision of feedback, extending previous research that has focused on the utility of scores assigned, to potential impact on learners. The results demonstrated lessened provision of feedback when additional demands were imposed despite the fact that both groups appeared equally concerned about candidates' performance given the GRS scores assigned. This study contributes to the growing body of research that has demonstrated impaired rater performance when demands associated with a rating task are high. While previous studies have demonstrated that the assessment *of* learning can be negatively affected, this study demonstrates that assessment *for* learning may be similarly disadvantaged if the rating task is poorly aligned with the inherent cognitive capacity of raters.

Performance impairments have been reported in a number of fields when demands associated with a task exceed an individual's cognitive capacity to manage task relevant features. For

example, in aviation, researchers have shown that high cognitive demands associated with flying can impair situational awareness (Alexander, 2000). Extensive research in instructional design settings consistently reveal learning impairments under similar conditions (Plass, Moreno et al. 2010) In health professions education, researchers are beginning to explore the role of mental workload in decision making (Byrne, 2012), on learning,(Van Merriënboer and Sweller, 2010), on the performance of standardized patients (Newlin-Canzone, et al. 2013), on clinical performance (Yurko, et al. 2010) and on rater based assessment,(Byrne, et al. 2014, Tavares and Eva 2014) each similarly finding impairments in performance under high mental workload conditions. Common among all these programs of research is the underlying principle that when tasks are dependent on humans and the task relevant and/or irrelevant but present information cumulatively exceeds the capacity of cognitive structures needed to complete the task, impairments can be expected. In this study, we demonstrated that task irrelevant information (i.e., not essential to assessing clinical competence) can impose demands that become additive with intrinsic demands and impair some aspects of rater performance.

One explanation for why the extraneous demands imposed in this study showed an effect that was not present in other manipulations used in previous work may be that raters were less able to apply avoidance strategies. In a previous study, when extraneous demands were imposed, participant interviews suggested that raters applied strategies to minimize cognitive work by doing things like prioritizing tasks and eliminating what they perceived to be extraneous features. When demands were high, raters would also idiosyncratically shrink their task (e.g., attending to only salient behaviours) in an attempt to restore balance or avoid cognitive work. In this study, raters in the immersed condition were obligated to "perform". Immersed raters were tasked with carrying out more than one role that was essential to the simulation, having to recite scripts or improvise, and forcing them to attend to different features the performance in order to complete their tasks effectively.

To our knowledge this is the first study to explore and demonstrate the impact of rating demands on formative feedback. Feedback has been recognized as essential in learning (Ende 1983), leading many to argue that, where possible, assessments of performance include formative feedback to avoid lost learning opportunities (Rushton, 2005; Saedon, et al. 2012; Kogan and Holmboe 2013). In this study, when rating demands were high, despite similar degrees of rater concern as indicated by rater leniency, we observed a reduced amount of specific task-centered feedback, reduced feedback related to dimensions of performance, reduced feedback targeting the individual and fewer directions and/or recommendations for improvement (i.e., impairments in 4 out 6 feedback types). We did not observe such differences in feedback related to the context or in a catch-all category of items not meeting any of the above criteria presumably because there were relatively few that fell into these categories. For educators, the implications associated with having additional extraneous tasks while engaging in formative assessments appear to be potentially costly for learners.

Moving forward, we agree with Kogan and Holmboe who argue that work based assessments will be optimized further, in terms of accuracy and provision of feedback, as additional research places the rater as the object of study (Kogan and Holmboe 2013). In this study while a statistically significant difference in the provision of feedback was observed it could be argued that the differences are small enough to have a negligible impact on student learning. Whether this is true or not cannot be determined from this study. However, this study does suggest the

presence of an effect that should be explored further. That is, when tasked with higher compared to lower demands, rater performance has now been seen to decline in several ways, suggesting a need to better understand the degree to which learners are affected. As such, rather than emphasizing mainly rater training, we suggest further identification of strategies to mitigate sources of challenge for raters, particularly with respect to intrinsic and extraneous load. The results of this study and of similar earlier research suggest both intrinsic and extraneous load may result in inappropriate rater idiosyncrasies or performance impairments when ignored. Strategies to address these issues may include separating the task of rating performance from additional simultaneous tasks (e.g. presentation of or responsibility for clinical cases). Such efforts will need to balanced against human resource availability and the need to ensure assessments of clinical performance that are context bound and not stripped of authenticity or meaningful construct representation.

Limitations

Conducting this study in the context of a real examination limited our ability to collect some information that would have been useful. For example, we were not able to test inter-rater reliability within experimental condition as was done in previous work even though this is the main form of reliability expected to be influenced by the cognitive demand placed on raters. Further, we did not have a method by which to assess the quality or impact of feedback. Future studies should ensure a criterion is available by which to draw such conclusions. Also, we used a deductive process for evaluating the provision of feedback, which may have limited our analysis of the data. Using a more flexible inductive / qualitative approach may reveal other findings and insights. Finally, checklists, may function as a cognitive aid given the additional detail, or create additional load given the number of items tends to be higher. Whether similar findings would occur with checklists (compared to a GRS) requires further study.

Conclusions

The results of this study contribute to a growing body of research exposing rater cognition as a real, but potentially modifiable limitation to the assessment of clinical competence. In many performance-based assessments, raters are responsible for additional roles extraneous to the rating task itself. A cognitive capacity model, wherein raters are recognized as having a finite capacity for holding task-relevant information in working memory would suggest that raters may have difficulty with such a complex task. In this study, rating clinical performance with additional extraneous tasks resulted in lessened provision of feedback.

Table 1: Feedback categories used to organize statements and their definitions.

Category	Definition	Example
Statement about specific behaviour or task performed	Statements may be positive or negative, involve commission or omission errors, prioritization or sequence issues, issues with timeliness or speed, or with integration or interpretation.	To ensure safety, when defibrillating a patient, be sure to have eyes on the patient and surrounding team members rather than the monitor.
General statement regarding a dimension of performance	Non-specific statements regarding situation awareness, history gathering, patient assessment, decision-making, resource utilization, communication and/or procedural skills.	Very little physical assessment done; a more thorough focused assessment is needed.
General statement regarding the individual	Summative statements regarding the individual's performance; judgments suggesting competence or incompetence.	Your communication was very effective.
General statement regarding the context	Statements made regarding the role, management or consideration of context or setting, both positive and negative.	Given the confined area, you did a good job of positioning your equipment where it was needed most.
Directions / Recommendations	Statements specifically targeting actions to be taken to improve, to sustain good behaviours or statements providing alternatives to consider in future.	More training is needed in the nuances of cardiac arrest guidelines; review those in more detail.
Encouraging reflection	Statements encouraging candidates to reflect on positive or negative elements or where there might be opportunity to invoke change or alternative actions	Think about how the bystanders could have been used more effectively.

Table 2: Demographics details by group.

Demographic Variable	Passive Raters	Immersed Raters
Gender	6 males 4 females	6 males 4 females
Years of clinical experience	9.6 years	10.1years
Familiarity with GRS	4.1/10	4.2/10
Primary Role (Clinician vs. Educator)	8 C / 2 E	6 C / 4 E
Previous rater training (yes?)	3/10	3/10

GRS = global rating scale. C = clinician. E = educator. Rater training unspecified for both groups.

Table 3: Mean global rating scores (95% CI) by dimension by group.

Dimension	Passive Rater Mean (95% CI)	Immersed Rater Mean (95% CI)	p value
Situation Awareness	4.9(4.7-5.2)	5.1(4.8-5.3)	.54
History Gathering	4.7(4.4-4.9)	4.8(4.5-5.0)	.58
Patient Assessment	4.5(4.2-4.7)	4.9(4.6-5.2)	<.05
Decision Making	4.5(4.2-4.8)	4.7(4.4-5.0)	.25
Resource Utilization	5.0(4.8-5.2)	5.1(4.8-5.3)	.77
Communication	5.3(5.0-5.6)	5.3(5.0-5.7)	.99
Procedural Skill	4.7(4.4-5.0)	5.0(4.7-5.2)	.19
Overall	4.7(4.4-5.0)	4.8(4.5-5.1)	.60

Table 4: Mean number of feedback statements (95% Confidence intervals) provided by raters in both groups as a function of feedback type.

Feedback Type	Passive Rater Mean (95% CI)	Immersed Rater Mean (95% CI)	p value
Specific Behaviour	7.0 (6.4-7.5)	5.2 (4.7-5.7)	p<.01
General Dimension	2.6 (1.7-3.4)	2.0 (1.4-2.6)	p<.01
General Individual	2.1 (1.9-2.3)	1.8 (1.6-2.1)	p<.05
General Context	1.0 (0.7-1.3)	0.8 (0.6-1.2)	p=.15
Provision of Direction	2.4 (2.0-2.7)	1.8 (1.6-2.1)	p<.05
Other	0.6 (0.4-0.9)	0.5 (0.3-0.7)	p=.08

Table 5: Individual variance components and percentage of total variance attributable to each facet analyzed separately by group.

	Passive		Immersed	
Effect	**Variance**	**% of Total Var.**	**Variance**	**% of Total Var.**
Candidate	0.12	6.7	0.08	4.9
Station(includes rater)	0.04	1.9	0.11	6.7
Items	0.08	4.2	0.02	1.5
Candidate x Station	0.59	32.0	0.64	38.3
Candidate x Items	0.03	1.7	0.01	1.1
Station x Items	0.24	8.0	0.11	6.5
Candidate x Station x Items	0.84	45.5	0.68	40.9
Internal Consistency	G=.08; G_n=.84		G=.44; Gn=.84	
Inter-station	G=.08; G_n=.61		G=.05; Gn=.49	

Var. = Variance

G Formulae:

Inter-station Reliability: [Var(c) + Var(ci)] / [Var(c) + Var(s) + Var(i) + Var(cs) + Var(ci) + Var(si) + Var(csi)]

Internal Consistency: [Var(c) + Var(cs)] / [Var(c) + Var(s) + Var(i) + Var(cs) + Var(ci) + Var(si) + Var(csi)]

G_n Formulae:

Inter-station Reliability: [Var(c) + Var(ci)/7] / [Var(c) + Var(s)/10 + Var(i)/7 + Var(cs)/10 + Var(ci)/7 + Var(si)/70 + Var(csi)/70]

Internal Consistency: [Var(c) + Var(cs)/10] / [Var(c) + Var(s)/10 + Var(i)/7 + Var(cs)/10 + Var(ci)/7 + Var(si)/70 + Var(csi)/70]

References

Alexander, A.L. (2000). Examining the relationship between mental workload and situation awareness in a simulated air combat task. DTIC Document.

Archer, J.C. (2009). State of the science in health professional education: effective feedback. *Medical Education*. *44*(1), 101-108.

Baddeley, A. (1992). Working memory. *Science. 255(*5044), 556-559

Bloch, R. & Norman, G. (2012). Generalizability theory for the perplexed: A practical introduction and guide: AMEE Guide No. 68. *Medical Teacher. 34*(11), 960-992.

Brennan, R.L., (2001). Generalizability theory. New York, NY: Springer Verlag.

Byrne, A., (2012). Mental workload as a key factor in clinical decision making. *Advances in Health Sciences Education. 18*(3), 537-545.

Byrne, A., Tweed, N., & Halligan, C. (2014). A pilot study of the mental workload of objective structured clinical examination examiners. *Medical Education*. *48*(3), 262-267.

Downing, S., & Yudkowsky, R. (2009). Assessment in Health Professions Education. New York, NY. Routledge.

Elo, S., Kyngäs, H. (2008). The qualitative content analysis process. *Journal of Advanced Nursing*. *62*(1), 107-115.

Ende, J., (1983). Feedback in clinical medical education. *JAMA. 250*(6), 777-781.

Govaerts, M., & van der Vleuten, C. (2013). Validity in work-based assessment: expanding our horizons. *Medical Education*. *47*(12), 1164-1174.

Govaerts, M., van de Wiel, M., Schuwirth, L., & van der Vleuten, C. (2011). Workplace-based assessment: raters' performance theories and constructs. *Advances in Health Sciences Education. 18*(3), 375-396.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*. *77*(1), 81-112.

Herbers, J., Noel, G., Cooper, G., Harvey, J., Pangaro, L., & Weaver, M. (1989). How accurate are faculty evaluations of clinical competence? *Journal of General Internal Medicine*. *4(*3), 202-208.

Kane, M. (2006). Validity. In *Educational Measurement. (4)* (pp. 17-64). Westport, CT: Praeger Pub Text.

Kane, M.T. (1992). The assessment of professional competence. *Evaluation & the Health Professions*. *15*(2): p. 163-182.

Kogan, J.R., & Holmboe, E. (2013). Realizing the Promise and Importance of Performance-Based Assessment. *Teaching and Learning in Medicine*. 25(sup1), S68-S74.

LaMantia, J., Rennie, W., Risucci, D., Cydulka, M., Spillane, S., Becher, J., & Kleinschmidt, M. (1999). Variability among Faculty in Evaluations of Residents Clinical Skills. *Academic Emergency Medicine*. *6*(1), 38-44.

Landy, F.J., Farr, J., (1980). Performance rating. *Psychological bulletin*. *87*(1), 72-107.

Miller, G. (1990). The assessment of clinical skills/competence/performance. *Academic Medicine*. *65*(9), S63-S67.

Morgeson, F., & Campion, M. (1997). Social and cognitive sources of potential inaccuracy in job analysis. *Journal of Applied Psychology*. *82*(5), 627-654.

Newlin-Canzone, E., Scerbo, M.W., Gliva-McConvey, G., & Wallace, A.M. (2013). The Cognitive Demands of Standardized Patients: Understanding Limitations in Attention and Working Memory With the Decoding of Nonverbal Behaviour During Improvisations. *Simulation in Healthcare. 8*(4), 207-214.

Noel, G., Herbers., J., Madlen, P., Caplow, G., Cooper, M., Louis, N., Pangaro, M., Havey, J. How well do internal medicine faculty members evaluate the clinical skills of residents? *Annals of Internal Medicine*, 1992. *117*(9), 757-765

Plass, J., R. Moreno, R., Brünken, R. Cognitive load theory. New York, NY: Cambridge University Press.

Rushton, A., (2005). Formative assessment: a key to deep learning? *Medical Teacher*. *27*(6), 509-513.

Saedon, H., Salleh, S., Balakrishnan, Imray, C., Saedon, M. (2012). The role of feedback in improving the effectiveness of workplace based assessments: a systematic review. *BMC Medical Education, 12*(1), 25.

Shute, V.J. (2008). Focus on formative feedback. *Review of Educational Research. 78*(1),153-189.

Tavares, W. & Eva, K.W. (2012). Exploring the impact of mental workload on rater-based assessments. *Advances in Health Sciences Education*. *18*(2), 291-303.

Tavares, W. Ginsburg, S., & Eva., K.W. (submitteed) Selecting and Simplyfying: Rater Behaviours and Performance when Considering Multiple Competencies.

Tavares, W. & Eva, K.W. (submitted). The Impact of Rating Demands on Rater Based Assessment of Clinical Competence.

Tavares, W., Boet, S., Theriault, R., Mallette, T., & Eva, K.W. (2012b). Global Rating Scale for the Assessment of Paramedic Clinical Competence. *Prehospital Emergency Care, 17*(1), 57-67.

Tavares, W., LeBlanc, V.R., Mausz, J., Sun, V., Eva, K.W. (2013). Simulation Based Assessment of Paramedics and Performance in Real Clinical Contexts. *Prehospital Emergency Care*. *17*(1), 57-67.

Van Merriënboer, J., & Sweller, J. (2010). Cognitive load theory in health professional education: design principles and strategies. *Medical Education*. *44*(1), 85-93.

Yeates, P., O'Neil, P., Mann, K., & Eva, K.W. (2013). Seeing the same thing differently. *Advances in Health Sciences Education*. *18*(3), 325-341.

Yurko, Y., Scerbo, M., Prebhu, A., Acker, C., & Stefanidi, D. (2010). Higher Mental Workload is Associated With Poorer Laparoscopic Performance as Measured by the NASA-TLX Tool. *Simulation in Healthcare. 5*(5): p. 267-271.

Chapter 7 – General Discussion and Conclusions

Introduction

The provision of safe, effective and patient centered health care is valued as fundamental in Canadian society and achieving these goals continues to be dependent on the individuals who are granted access to and serve within the health professions. This places considerable importance on the education trainees receive but also on the assessment processes used to make decisions regarding achievement in domains of competence. As the concept of "competence" in many health professions continues to become increasingly complex due to a greater emphasis on the integration of multiple competences (Frank, et al. 2010) and a broadening of focus (Frank 2005), a greater emphasis has been placed on encouraging assessment strategies that involve direct observation of trainees in clinical practice settings (Eva and Hodges 2012; Hodges 2013). However such judgments require complex cognitive processes and raters have been identified as fallible in these contexts (Tavares and Eva 2012). Attempts to mitigate challenges attributable to raters have been insufficient with limited or incomplete explanations as to why this task is so difficult. The work reported in this thesis examined the role of raters in the assessment process and describes, in particular, the impact of rating demands on rater performance. I hypothesized that as rating demands increase in rater based assessments of clinical competence, indicators of rating quality suffer as a result of raters engaging in cognitive behaviours that result in idiosyncrasies related to satisfying rather than optimizing rater performance. Next I summarize the main findings from the research conducted to explore this hypothesis, discuss their implications for health professions education and outline future directions.

Summary of Findings

A targeted review of multiple literatures revealed many reasons to believe that rater based assessments in health professions education may be fundamentally influenced by rating demands that are imposed (and usually ignored) in assessments of clinical competence(Tavares and Eva 2012). That is, the complexity of behaviours that can exist within domains of competence, the sheer number of behaviours that need to be observed and processed (including their interactions), the influence of clinical or assessment contexts, trainee characteristics, the requirement to retrieve performance standards in real time, etc. may result in excessively high mental workload and ultimately induce rater idiosyncrasies. Theories of attention, (Kahneman 1973) perception, (Lavie 1995) working memory, (Baddeley 1992) information processing, (Wickens and Carswell 2006) multiple resource,(Wickens 2008) and cognitive load, (Sweller 1988) suggest that the necessary cognitive systems are capacity limited. Therefore, when strategies to minimize or reduce mental workload are unavailable, ineffective or ignored, or as the information becomes increasingly complex, the potential for overload increases and can lead to performance impairments, avoidance and/or simplifying strategies. The extent to which the task assigned to raters aligns with their cognitive and perceptual capacities determines the extent to which reliance on human judgment threatens assessment quality. The multifaceted nature of the construct of clinical performance, the reliance on rater judgment and the unique and dynamic settings in which clinical assessments are expected to take place, suggested the need for field specific studies of the alignment between mental workload and cognitive capacity in health professions education.

Before exploring these concepts, we conducted a scale development study to ensure that the metric used was conceptually meaningful in the study context. This development process resulted in a 7 dimension global rating scale (GRS) built using the conventions of scale development including construct representation. When we tested the scale using 2 raters evaluating a total of 80 videos including trainees, entry to practice level candidates and experienced clinicians, inter-rater reliability, intra-rater reliability and internal consistency was moderate to high (see chapter 3 for details) and scores increased with level of education/experience. With a rating scale in place and using cognitive capacity as a conceptual framework we conducted three experimental studies to test two primary hypotheses: (a) as rating demands increase in rater based assessments of clinical competence, indicators of rating quality or rater performances would decline; (b) when rating demands exceed cognitive resources, raters would engage in cognitive behaviours that result in idiosyncrasies aimed at satisfying rather than optimizing performance.

The second study titled "The Impact of Rating Demands on Rater Based Assessments of Clinical Competence" (see chapter 4) tested the first hypothesis by manipulating intrinsic and extraneous load in a 2x2 factorial design. Factor A represented intrinsic load or essential processing (tasks directly related to rating a trainee) by having raters consider more or less dimensions simultaneously while rating three independent clinical performances. Factor B represented extraneous load (tasks not directly related to rating a trainee) by manipulating the presence or absence of instructions requiring participants to monitor the patient's condition and suggest alternatives in clinical care that might be needed.

As predicted, as the number of dimensions to be considered increased from 2 to 7, indicators of rating quality including the ability to detect dimension relevant behaviours, the ability to discriminate between levels of performance and inter-rater reliability declined. When reliabilities were calculated by averaging over multiple raters, ceiling effects were observed. We expected to observe further impairments in rater performance when extraneous load was induced, but did not observe such significant effects. We speculated that participants in all groups might have integrated the additional tasks we selected into their cognitive activity spontaneously because it is difficult for clinicians to not attend to clinical problems, thereby eliminating the intended difference between groups. Still the results of this study uncovered a potential link between rating quality and rating task that had not been identified previously and raised important challenges for the assessment community to consider.

The purpose of the third study, titled "Selecting and Simplifying: Rater Behaviour and Performance when Considering Multiple Competencies" was to: (a) further explore the effect of rating demands on rater performance in the assessment of clinical competence; (b) test our second hypothesis that when rating demands exceed cognitive resources raters engage in cognitive behaviours that result in idiosyncrasies; (c) seek an explanation for the conflicting results observed in the first study; and (d) determine if content expertise could serve as a mitigating factor (see chapter 5). This led to us to conduct a parallel mixed methods study to collect both quantitative and qualitative data. For the quantitative portion of the study we replicated our first experimental study but with important differences. First, we conducted a similar 2x2 factorial design using the same manipulation for intrinsic load (i.e., 7 vs. 2 dimensions) but applied a different manipulation to induce extraneous load. Rather than assign additional tasks that made it difficult to confirm if they were undertaken, we assigned half of all

participants to additionally rate the performance of 2 standardized health care professionals who were observable in the three rating videos. Second, we used expert raters rather than novice raters. Expert raters are more likely to have cognitive resource sparing schemas, thereby allowing us to extend our proof of concept and test the robustness of our previous study results. Third, given the large percentage of total variance attributed to our facet of differentiation (or subject / video) in our first experimental study we used a different set of videos that were expected to be more homogenous to determine if the effect persisted. Finally, the qualitative portion of the study involved interviewing participants in the high load group to explore strategies they engaged while completing the rating task under those conditions.

The quantitative portion of this study supported our primary hypothesis and replicated our earlier results on two levels. First, as the number of dimensions to be considered was greater, the ability to identify dimension relevant behaviours and the ability to differentiate between candidates (i.e., reliability) declined. In this study the percentage of total variance attributed to our facet of differentiation (video) was lower relative to the preceding study, suggesting that the videos were in fact more homogeneous set. A review of the differences between means in our first and second experimental studies further suggests that to be the case as the difference between means for each of the two dimensions was larger in the first study compared to the second. Nonetheless, the key pattern of results (greater inter-rater reliability in the 2D condition relative to 7D) remained the same, supporting the robustness of this finding. It should be noted, however, that the greater homogeneity relative to the preceding study made it such that the reliabilities of single ratings demonstrated a floor effect and that, as a result, the effect of number of dimensions was only strongly evident when D-studies were performed to examine the reliability of the average calculated across multiple raters. Having to consider multiple dimensions or competencies simultaneously appears to consistently increase mental workload to the point of exceeding inherent cognitive resources and impairing performance on intended goals regardless of raters' experience or level of content knowledge. Second, we again failed to observe a meaningful effect as a result of the additional tasks we imposed. When we explored the point-in-time cognitive strategies raters engaged when presented with high rating demands, we identified some behaviour that help explain our consistent findings for both the impaired rating performance and lack of effect associated with the extraneous tasks.

Our qualitative analysis of post-task interview data in this study revealed two main strategies raters adopted when presented with mental workload. The first involved a process of "selection" whereby raters would reduce demands themselves by focusing on fewer dimensions when all could not be considered simultaneously. Which dimensions participants chose to focus on varied idiosyncratically and was difficult for raters to explain. The second main strategy involved a process of "simplification". Here participants would find and apply whatever behaviours they could to simplify the rating process in an effort to minimize cognitive work. For example, raters chose or found themselves focusing on only salient behaviours, or exclusively on positive or negative behaviours, and/or eliminating from attention anything they perceived to be extraneous, including the additional task of considering the standardized healthcare professionals observable in our stimuli. These behaviours shed light on potential causal factors for both of the consistent findings we observed: (1) those with more competencies to consider performed poorer than those with fewer; and (2) manipulating load using extraneous variables failed to have a similar influence. Participants simply managed their own load in ways that may have been helpful to minimize cognitive overload but also resulted in idiosyncrasies and performance that

significantly differed from expectations (e.g., that each dimension be considered as equally relevant). With the number of dimensions to be considered simultaneously consistently impairing rating performance and the identification of cognitive behaviours that explained much of our results, we sought to push the extraneous load manipulation further to determine if an extraneous task that was nonetheless essential to the assessment would result in lessened rating quality.

Given the consistent effect of a relatively minimal intrinsic manipulation and lack of effect of extraneous variables on rating quality, the purpose of the next study titled "Passive vs. Immersed: Rater Performance under Different Load Conditions" was to engage raters with a more essential extraneous load manipulation (see chapter 6). In this experimental study participants were randomly assigned either the role of "passive" or "immersed" rater in an existing OSCE. Participants assigned to the "passive" condition were required to assess candidates' performances using a 7 dimension GRS with no additional tasks. In contrast, participants in the "immersed" condition were also required to assess candidate performance using the same 7 dimension GRS but with the additional task of playing a role in the simulation (i.e., a concerned family member). This extraneous task is not one that participants could simply ignore because it required them to respond to candidate inquiries. Raters were paired such that each station included both a passive and immersed rater. As assessments can be both formative and summative, we also extended the methodology by exploring the impact of rating demands on the provision of feedback. In addition to rating the candidates' performance, both groups were required to provide formative feedback directly on the GRS. This manipulation revealed significant differences in terms of the amount of feedback delivered with lesser amounts of feedback being observed in the immersed condition despite ratings of performance being equivalent between both groups.

In summary the results of a targeted review suggests, and three experimental studies confirm, that rating demands can be misaligned with inherent human cognitive capacity and result in rater idiosyncrasies and performance declines on indicators of rating quality. This lack of alignment may lead raters to engage in cognitive processes designed to manage their mental workload, resulting in variation in behaviours observed and ultimately in what behaviours are processed for the sake of rating performance. Under high load conditions, raters appear to manage load in attempts to satisfy rather than optimize performance(Simon 1972) and in doing so may eliminate what they perceive (even if incorrect) to be extraneous elements of their tasks. This explains both the idiosyncrasies and performance impairments observed as well as the difficulty inducing an impactful extraneous load if the extraneous task is non-essential. To my knowledge, this is the first set of studies to explore these particular theories in the assessment of clinical competence and the cognitive behaviours raters engage when presented with high mental workload demands. Next, we discuss practical and theoretical implications for health professions educators and researchers, limitations in our research and future directions.

Implications for Health Professions Education

In health professions education, optimizing rater based assessment of clinical competence where direct observation of candidates is applied, will require that educators and researchers give attention to an interaction that exists between rater cognition, assessment frameworks and supporting systems. See figure 1 for an illustration of this model, which is intended to represent a supplemented view of levels 3 (shows how) and 4 (does) of Miller's pyramid (Miller 1990).

First, rater cognition refers to rater characteristics and recognizes raters as integral but also as cognitive filters in the process of evaluating performance. A number of perspectives can apply. For instance, impression formation(Wood 2013), social judgments (Gingerich et al. 2011) assessor biases (Yeates, et al. 2013) and a rater's own clinical skills / expertise (Govaerts, Schuwirth et al., Kogan, Hess et al. 2010) may each influence the rating process in different ways. Second, assessment frameworks refer to decisions in assessment designs that inform implementation plans such as prioritizing formative vs. summative assessment goals, the development and/or implementation of one scale type over another, isolated task vs. integrated performance, simulation based vs. workplace based sampling, etc. Third, supporting systems include processes that may be in place to support successful assessment such as explicit institutional policies, alignment with the curriculum (explicit and hidden), validity frameworks, rater training etc. Given my program of research, I will focus on the interaction between the influence of mental demands (rater cognition), scale development (assessment framework) and rater training (supporting systems).

Focusing first on the rater cognition side of the pyramid, examining the role of rater cognition reveals that this is inherently a complex task. One of the challenges associated with levels 3 and 4 is that clinical competence cannot be measured directly, but rather must be derived using rater inferences and judgments. Both are dependent on the behaviours raters attend to and process while observing candidates perform in response to clinical challenges. Challenging this process, raters must observe and process reams of information often involving multiple micro or meta-competencies simultaneously, while also taking into account their interactions and the influence of the context (e.g., clinical case, social, physical and environmental characteristics, candidate characteristics, etc.) in which they are occurring. All of this must then be meaningfully translated into some form of categorical judgment or intelligible narrative to summarize the observations or form meaningful feedback either during point-in-time observations of performance or immediately following. At the information acquisition and processing stages of this process, structures needed to perform these tasks are capacity limited and, as we have demonstrated, when demands exceed resources, impairments in performance can be expected, threatening the overall process.

The importance or necessity of considering rating demands in rater based assessments is analogous to instructional design research which has demonstrated that learning becomes impaired when educational strategies (instead of rating demands) exceed cognitive resources (Plass, et al. 2010). Instructional design research is based on similar cognitive science, which suggests that many of the cognitive structures needed for learning are capacity limited and only a limited amount of information can be selected, organized, integrated or processed at any one time (Chandler and Sweller 1991, Baddeley 1992; Mayer 1999, Mayer and Moreno 2003). Successful implementation of educational strategies involves an appreciation for the ever-present risk of overloading cognitive resources.

Researchers have developed a number of best practices intended to ensure alignment between a learner's cognitive capacity and the educational material presented to avoid placing learners are at risk of becoming overloaded (Mayer and Moreno 2003; Van Merriënboer, et al. 2003; Van Merriënboer and Sweller, 2010) that may translate usefully to facilitating improved rater performance. For example, "segmenting" involves breaking the entire learning task into manageable chunks or allowing learners to intellectually digest some information before moving

on to the next. Importantly, the effect can be strengthened when the learner has control over the segmenting of information suggesting the load can be managed internally to some extent (Mayer and Chandler 2001). Another strategy referred to as "pre-training" involves providing learners with information and an opportunity to learn necessary component parts associated with the eventual learning task, in advance (Mayer and Moreno 2003). Common among these and other strategies are an appreciation for (a) the inherent human capacity-limited structures needed to engage in learning (e.g., attention, working memory), (b) the impairments associated with overloading learners and (c) the value and effectiveness of minimizing intrinsic load / essential processing demands or other types of load (i.e., avoiding the threshold of capacity) in optimizing performance. Strategies to do so have mainly been "controlled" by the instructional designer in advance of the learning session. While this is clearly an appropriate and effective strategy, the context of work-based assessments and the push to consider the integration of multiple competencies simultaneously can limit the extent to which such control can be achieved in many rater-based assessment contexts. As such, just as learners are left to manage excessive load internally when instructional designs fail or best practices are ignored, raters are similarly left to engage in internal strategies to manage excessive load when rating systems imposed on them ignore working memory capacity. Internal management of cognitive load thereby provides a rich field of research to explore in both instructional design and rater-based assessments. Using these foundations, the model illustrated in Figure 1, and keeping rating demands and a cognitive capacity framework in mind, we describe available strategies that might advance best practices in rater-based assessment.

A second side of the pyramid is labeled "assessment framework (e.g., scale development)" and it is at this point where an interaction between elements (or sides of the pyramid) becomes evident. Recognizing many of the rater-based challenges that persist, educators and researchers often develop rating tools using well established rules to assist raters in this complicated process (Streiner and Norman, 2008). Validity requirements for example, include ensuring the scale completely represents the construct of interest (Downing 2003, Downing and Haladyna 2004). Currently, literature informing scale development processes, however, offer very little consideration, if any, to how processes associated with ensuring construct representation impacts or interacts with rater cognition / rating demands. For example, attempts to improve internal consistency of scales (a form or reliability) often include increasing the number of items on the scale; a recommendation clearly in conflict with the findings of our research. Rating tools will continue to be valuable in assessment of competence, however, what may need to be considered are modifications to the development and implementation process when recommendations and rating demands are at odds.

Researchers are beginning to explore the benefits of considering raters in the scale development and implementation process. For example, Crossley and Jolly argue that the cognitive characteristics of raters have a significant influence on rating processes and that where possible features of rating tools should align with rater conceptions (Crossley and Jolly 2012). Consistent with this notion, participants in our research indicated difficulty in aligning behaviours with the dimensions included on our scale (i.e., matching behaviours observed with dimension definitions or expectations), which may have forced them to work with the information for longer periods of time, thereby increasing memory load. Our findings suggest that researchers should further study the effect of modifying scales to avoid greater degrees of intrinsic / essential processing load. For example, while some scale dimensions may require observation of behaviours periodically (e.g.,

procedural skills in a clinic visit) others may be continuous and overarching (e.g., situation awareness). Perhaps it is necessary, as a result, to tailor the dimension of focus to the specific moment of assessment rather than trying to do all things in all contexts. Clearly any modifications to the scales or scale development processes have validity and process implications that will need to be explored and resolved. In sum, taking into account rater cognition (specifically rating demands) will require modifications to scale development guidelines, new theoretical or practical views about assessment, and/or efforts aimed at striking a balance between what the assessment community can expect from raters and existing validity frameworks. The important point as demonstrated by this body of research and as illustrated by figure 1, is that scale development now be recognized as attached or linked to rater cognition in the overall process of direct observation. Emphasizing one without the other may simply be incomplete or flawed.

The third side of the pyramid is labeled "supportive systems (e.g., rater training)" and interacts with both the "scale development" and "rater cognition" sides of the pyramid. Rater training aims to improve rater performance by developing the necessary knowledge, skills and attitudes to accurately evaluate demonstrated skills and competencies"(Feldman, Lazzara et al. 2012) which traditionally has not included any explicit consideration for management of rating demands. Management of load in general is one of the fundamental tenets in instructional design research (Van Merriënboer and Sweller 2010) but is typically applied by the instructor / educator. For example, eliminating redundancy, using simple to complex strategies or worked examples, segmenting, signaling or pre-training are all available strategies. See Mayer and Moreno (2003) and Van Merrienboer and Sweller (2010) for a comprehensive overview of these and other instructor-applied strategies (Mayer and Moreno 2003, Van Merriënboer and Sweller 2010). However, when instructor-applied principles are ignored or not included, learners are left to struggle. Recently, researchers have revealed that teaching learners to manage their own load is not only possible but also effective (Roodenrys, et al. 2012). In an experimental study, when learners were given guidance on how to deal with materials inducing split attention (i.e., educationally relevant materials separated by time or space) this group of learners not only compared favorably with learners in the instructor-corrected materials, but were able to transfer their strategies to novel contexts (Roodenrys, et al. 2012). Applied to rater-based assessments, when raters are tasked with rating stimuli that exceed resources and assessment designs fail to reduce rating demands effectively or enough, teaching raters how to manage their own load may be beneficial.

Two forms of rater training with the greatest evidence of effectiveness (at least in the performance appraisal literature) include performance dimension and frame of reference (FOR) training. (Roch, et al. 2011) In performance dimension training raters are familiarized with the dimensions on which performance will be evaluated (Woehr and Huffcutt 1994), whereas FOR training provides raters with common conceptualizations or standards of performance relating to those dimensions. (Woehr and Huffcutt 1994, Roch, et al. 2011) It is possible, even likely, that rater training of this kind results in shared mental models that may promote inter-rater reliability. It is also possible that some schema development occurs that may ultimately lead to cognitive efficiencies and the ability to manage rating demands more effectively. However, empirical findings in support of rater training in health professions education have been equivocal. The results of our research suggest that the complexity associated with clinical competence and/or the failure to account for rating demands may be one reason why.

The qualitative results of our "Selecting and Simplifying" study were the first to suggest raters are actively managing their own load even though that management might be ineffective. When intrinsic load was high, raters engaged in a process similar to what instructional design researchers describe as "weeding"; a process of eliminating interesting but extraneous stimuli. Again, in instructional design settings, this is typically instructor applied (Mayer and Moreno 2003). For example, an embellished narrated animation, one that includes background music or other irrelevant information increases incidental processing or extraneous demands and has demonstrated impairments in problem solving or transfer compared to groups where this extraneous material was omitted (Mayer, et al. 2001). Participants in our research described behaviours similar to "weeding", doing this on their own with what they perceived to be extraneous information but also in making attempts to reduce intrinsic load by selecting fewer dimensions and/or easier elements / behaviours (e.g., attending only to salient behaviours). Modifications to rater training might include training raters to identify and appropriately eliminate behaviours, events or stimuli that are extraneous and induce incidental processing. Alternatively, training raters to become aware of strategies to reduce intrinsic load and working within those boundaries as dictated by their capacity in a given context (e.g., assigned scale, clinical case / encounter, contextual forces etc.) rather than forcing flawed processes (e.g., trying to consider all dimensions equally) may prove beneficial. While additional research is clearly needed in this area, rater training may continue to be limited in effectiveness until it can include strategies aimed at teaching raters how to manage load and/or work truly within their own capacity.

We propose then, given the complexity with which raters must contend, that rater based assessments of clinical competence consider all sides of the supplemented pyramid interactively rather than in isolation or in sequence. The degree to which one is emphasized over the other or the lengths to which one must go to achieve stated goals in one over the other will be context dependent. The guiding principle is, however, that rating demands be managed externally and/or internally to a point where an alignment between rating demands and inherent human cognitive architecture is achieved. The combination of interventions will ultimately, again, be context dependent, but must be considered.

Limitations

There are some potential limitations with the work presented within this dissertation that might affect generalizability. First, we used one type of global rating scale throughout. The decision to use a global rating scale in this study was based on the current direction of the assessment literature, which has increasingly supported the use of GRSs and rater judgment over checklists in similar contexts (Norman, Vleuten et al. 1991, Vleuten 1996, Crossley, Humphris et al. 2002).The unique characteristics of this particular rating scale may have resulted in degrees of mental workload that may be different with scales using different dimensions, structure or features. However, the scale was developed using widely accepted standards and is not drastically dissimilar to other rating tools used in health professions education. (Norcini, Blank et al. 1995, Kim, Neilipovitz et al. 2006) Other types of rating tools (e.g., task specific checklists) may result in more or less mental workload and performance results. Second, we used a simulation-based context involving a complex clinical case. The number of competencies and speed at which they occurred in our experimental studies may also combine to create challenges for raters that may not exist in other contexts. Also real contextual forces (e.g., true concerns for

patient safety) may not have been replicated effectively. Therefore generalizing results to other types of cases and settings may be limited. Future research will need to determine in what way these meaningful factors affect rating demands and ultimately rater performance. Similarly, our procedures for 2 of the 3 experimental studies required raters to verbally identify dimension relevant behaviours before forming judgments. This may have affected their natural assessment tendencies. While this manipulation was held constant when used to avoid confounding our results, overall performance may have been impacted. However, as judgments are ultimately based on the behaviours observed, the procedure provided an indication of what informed those judgments. Finally, the qualitative portion of this dissertation involved using verbal reports as data. Our interviews were conducted post-task and, unlike think alouds, encouraged raters to reflect on their strategies when completing their rating tasks. This procedure is inherently limited as it assumes raters are capable of reflecting accurately and are cognitively aware of their cognitive behaviours in retrospect (Nisbett and Wilson 2005). Future research will need to identify other procedures for identifying cognitive behaviours in order to better understand rater performance.

Future Directions

There are a number of research pathways and questions that have been raised and are left unanswered given this new program of research. We have supplemented a well-established model in the assessment of clinical competence and contributed to emerging discourse by suggesting a caution and direction in assessment when considering rater judgment. A number of research questions and future directions have been suggested throughout and rather than repeating them here I have chosen to focus on three areas of research that, if carried through, will extend the work presented here, further our understanding of cognitive behaviours raters engage in the context of assessment, and provide an opportunity to continue to contribute to emerging discourses and broadening perspectives regarding assessments.

Focusing first on extending the results of this dissertation, future research will need to explore the "tipping point" at which impairment occurs when considering multiple competencies and what factors other than number of dimensions lead to excessive mental workload. Knowing the threshold at which raters' transition from a point of alignment between rating demands and available resources to one that results in poor alignment or overload has important theoretical and practical implications. Rather than focusing solely on quantity, it would also be helpful to explore the characteristics of load inducing variables (e.g., is assessing communication more cognitively demanding than procedural or technical skills) across contexts to determine their generalizability. Similarly, understanding what factors contribute to (or reduce) mental workload will also inform educators on process recommendations. Extending this work beyond simulation-based settings where a number of real contextual forces become influential, is also needed. Finally, establishing strategies that can shift the "tipping point" or threshold to foster resiliency through better management of rating demands internally (i.e., within the rater) and externally (i.e., eliminating extraneous load) will be especially informative.

Second, in our review of the literature we explored medical decision-making and reasoning to understand individual behaviours in the way clinicians respond to cognitive work. Not discussed in great detail was research exploring dual process theory where clinicians engage system 1 (rapid, effortless, automatic) and system 2 (slow, effortful, conscious) thinking styles as appropriate or necessary for a given context (Eva 2005). A common assumption in this research

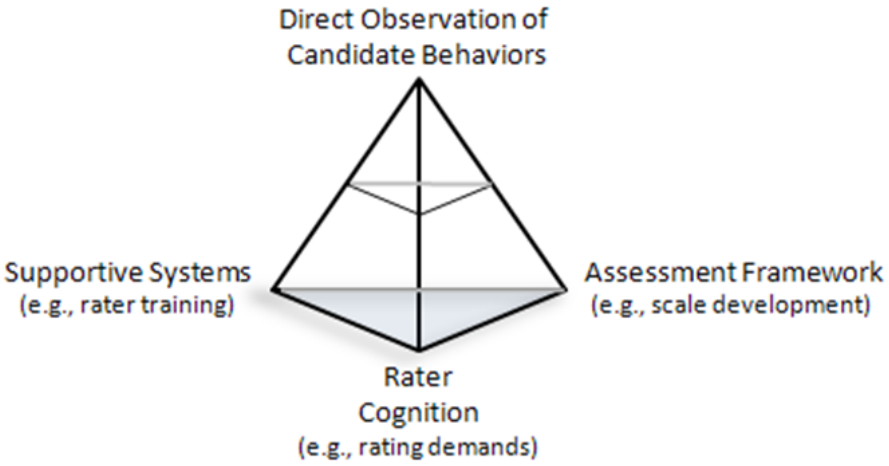
is that system 1 uses comparatively less cognitive resources compared to system 2. In response to our exploration of rating demands on rater performance, researchers have drawn on this literature to explore rating processes and performance from this perspective (Wood 2013). Some researchers suggest dual process and related theories may not fully explain clinical decision-making and suggest instead, the value of exploring mental workload (Byrne 2012). Other areas of clinical reasoning may also inform rater cognition research. For example, we know that cognitive errors in medical decision-making are often linked to faulty data collection, interpretation, reasoning and/or biases and other mental shortcuts, and that much of this has been attributed to limitations in human processing (Graber, et al. 2002). Merging these two programs of research may be fruitful. For instance, exploring the degree to which one system is used over the other, when, under what circumstances and what the implications are may further our understanding and recommendations toward assessment frameworks (e.g., reliance on rater judgment) and supportive systems (including what we may and may not reasonably expect to accomplish).

Finally, when rater based assessments are viewed using a cognitive capacity perspective (i.e., rater cognition), in addition to informing assessment frameworks and supporting systems, future research can also contribute to a recent and growing discourse involving the use of rater judgment in general. Hidden in many initial attempts to mitigate rater based issues, was the direct or indirect perceptions that raters could be viewed as objective measurement devices accurately transferring information from behaviours, to rating tools to decisions. However, more recently drawing analogies to decision making in medicine where judgment is recognized as important (but not infallible), researchers have argued for a greater reliance on rater judgment (Hodges 2013). Our research suggests that embracing inference and judgment in assessment contexts requires an appreciation for role of the rater as a cognitive filter in the process. That is, challenges, opportunities and/or effective assessment strategies associated with greater reliance on "subjective" rater based assessments, may be strengthened by considering and understanding the impact of rating demands on rater inferences and judgments.

Throughout the research presented here we have suggested that rater idiosyncrasies are viewed as error or noise in the assessment process and have made recommendations on how the ratio of signal to noise might be improved. Consistent with previous research (Yeates, et al. 2012) we have demonstrated significant idiosyncrasies in what raters attend to but tried to explain these findings using a limited capacity framework. As such, discordance between rating demands and cognitive capacity might help to explain why different raters can have equally valid yet incommensurate perspectives on an individuals' performance. If true, this model might be helpful in health professions education to explore when differences in rater judgment should be considered to represent noise or signal.

In summary, in addition to the research agendas outlined when considering implications for health professions education (e.g., innovative rater training strategies, modifications to scale development processes including implications to validity frameworks) identifying a tipping point, further exploring cognitive behaviours and studying the degree to which rater idiosyncrasies are better viewed as signal or noise will advance the research included in this dissertation.

Figure 1: Assessment pyramid demonstrating the interacting elements that should be considered in the assessment of competence.



References:

Albanese, A. (2000). Challenges in using rater judgements in medical education. *Journal of Evaluation in Clinical Practice*. *6*(3), 305-319.

Alexander, A. L. (2002). Examining the relationship between mental workload and situation awareness in a simulated air combat task. *DTIC Document*.

Archer, J. C. (2009). State of the science in health professional education: effective feedback. *Medical Education*. *44*(1), 101-108.

Baddeley, A. (1992). Working memory. *Science. 255(*5044), 556-559.

Bloch, R., & Norman, G. (2012) Generalizability theory for the perplexed: A practical introduction and guide: AMEE Guide No. 68. *Medical Teacher*. *34*(11), 960-992.

Bloom, B. (2002). Crossing the quality chasm: A new health system for the 21st century. *JAMA. 287*(5), 646-647

Bogo, M., Regehr, C., Woodford, M., Hughes, J., Power R., & Regehr, G. (2006). Beyond Competencies: Field Instructors Descriptions of Student Performance". *Journal of Social Work Education*. *42*(3), 579-594.

Bogo, M., Regehr, C., Power R., & Regehr, G. (2007). When Values Collide. *The Clinical Supervisor. 26*(1), 99-117.

Brennan, R. L. (2001). Generalizability Theory. New York, NY: Springer Verlag.

Byrne, A. (2012). Mental workload as a key factor in clinical decision making. *Advances in Health Sciences Education*. *18*(3), 537-545.

Byrne, A., Tweed, N., & Halligan, C. (2014). A pilot study of the mental workload of objective structured clinical examination examiners. *Medical Education. 48*(3), 262-267.

Carraccio, C. L., & Englander, R. (2013). From Flexner to Competencies: Reflections on a Decade and the Journey Ahead. *Academic Medicine. 77*(5): 361-367.

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction* . *8(*4), 293-332.

Charmaz, K. (2006). Constructing grounded theory: A practical guide through qualitative analysis. London: Sage Publications Ltd.

Cook, D. A., Beckman, T.J., Mandrekar J.N., & Pankratz, V.S. (2010). Internal structure of mini-CEX scores for internal medicine residents: factor analysis and generalizability. *Advances in Health Sciences Education*. *15*(5): 633-645.

Cook, D., Dupras, D., Beckman, T., Thomas, K., & Pankratz, V.S. (2009). Effect of rater training on reliability and accuracy of mini-CEX scores: A randomized, controlled trial. *Journal of general internal medicine. 24*(1), 74-79.

Creswell, J. W., Clark, V.L.P. (2007). Designing and conducting mixed methods research, Thousand Oaks, CA: Wiley Online Library.

Croskerry, P. (2003). The importance of cognitive errors in diagnosis and strategies to minimize them. *Academic Medicine. 78*(8), 775-780.

Crossley, J., & Jolly, B. (2012). Making sense of work-based assessment: ask the right questions, in the right way, about the right things, of the right people. *Medical Education*. *46*(1), 28-37.

Crossley, J., Humphris, G., & Jolly, B. (2002). Assessing health professionals. *Medical Education. 36*(9), 800-804.

Crossley, J., Johnson, G., Booth, J., & Wade, W. (2011). Good questions, good answers: construct alignment improves the performance of workplace based assessment scales. *Medical Education. 45*(6), 560-569.

DeNisi, A. (1996). A cognitive approach to performance appraisal: A program of research, New York, NY: Routledge.

Downing, S. (2003). Validity: on the meaningful interpretation of assessment data. *Medical Education. 37*(9), 830-837.

Downing, S. (2004). Reliability: on the reproducibility of assessment data. *Medical Education. 38*(9), 1006-1012.

Downing, S. (2005). Threats to the validity of clinical teaching assessments: What about rater error? *Medical Education*. *39*(4), 353-355.

Downing, S., Yudkowsky, R. (2009). Assessment in Health Professions Education,. New York, NY: Taylor & Francis.

Downing, S., & Haladyna, T. (2004). Validity threats: overcoming interference with proposed interpretations of assessment data. *Medical Education. 38*(3), 327-333.

Elo, S., & Kyngäs, H. (2008). The qualitative content analysis process. *Journal of Advanced Nursing*. *62*(1), 107-115.

Ende, J. (1983). Feedback in clinical medical education. *JAMA* 250(6), 777-781.

Epstein, R., Hundert, E. (2002). Defining and assessing professional competence. *JAMA. 287*(2): 226-235.

Eva, K.W (2005). What every teacher needs to know about clinical reasoning. *Medical Education. 39*(1): 98-106.

Eva, K.W (2010). Assessment Strategies in Medical Education. *Medical Education: State of the Art.* R. Salerno-Kennedy and S. O'Flynn. (pp. 93-106). Halifax, NS: Nova Scotia Publishers.

Eva, K. W., & Hodges, B.D. (2012). "Scylla or Charybdis? Can we navigate between objectification and judgement in assessment?" *Medical Education*. 46(9), 914-919.

Feldman, M., Lazzara, E.H., Vanderbilt, A.A., & Diaz Granados, D. (2012). Rater training to support high-stakes simulation-based assessments. *Journal of Continuing Education in the Health Professions.* 32(4), 279-286.

Frank, J. (2005). The CanMEDS 2005 physician competency framework. Better standards. Better physicians. Better care. Ottawa: *The Royal College of Physicians and Surgeons of Canada.*

Frank, J. R., Snell, L.S., Cate, O.T., Holmboe, E.S., Carraccio, C., Swing, S.R., Harris, P., Glasgow, N.J., Campbell, C., & Dath, D. (2010). Competency-based medical education: theory to practice. *Medical Teacher.* 32(8), 638-645.

Frank, J. R., Mungroo, R., Ahmad, Y., Wang, M., De Rossi, S., & Horsley, T. (2010). Toward a definition of competency-based education in medicine: a systematic review of published definitions. *Medical Teacher*. 32(8): 631-637.

Gingerich, A., Regehr, G., & Eva, K.W. (2011). Rater-Based Assessments as Social Judgments: Rethinking the Etiology of Rater Errors. *Academic Medicine.* 86(10), S1-S7.

Ginsburg, S., McIlroy, J., Oulanova, O., Eva, K.W., & Regehr, G. (2010). Toward Authentic Clinical Evaluation: Pitfalls in the Pursuit of Competency." *Academic Medicine*. 85(5): 780-786.

Gopher, D., Braune, R. (1984). On the psychophysics of workload: Why bother with subjective measures? *Human Factors*. 26: 519-532.

Govaerts, M., Van der Vleuten, C.P. (2013). Validity in work-based assessment: expanding our horizons. *Medical Education.* 47(12): 1164-1174.

Govaerts, M., Schuwirth, L.W., Van der Vleuten, C.P. & Muijtjens, A.M.M., Workplace-based assessment: effects of rater expertise. *Advances in Health Sciences Education*. 16(2): 151-165.

Govaerts, M., van de Wiel, M., Schuwirth, L, Van der Vleuten, C.P. (2011). Workplace-based assessment: raters' performance theories and constructs. *Advances in Health Sciences Education.*18(3), 375-396.

Govaerts, M., Van der Vleuten, C.P., Schuwirth, L.W., & Muijtjens, A. (2007). Broadening perspectives on clinical performance assessment: rethinking the nature of in-training assessment. *Advances in Health Sciences Education*. 12(2), 239-260.

Graber, M. (2009). Educational strategies to reduce diagnostic error: can you teach this stuff? *Advances in Health Sciences Education*. 14(1), 63-69.

Graber, M., Gordon, R., & Franklin, N. (2002). Reducing diagnostic errors in medicine: what's the goal? *Academic Medicine*. 77(10), 981-992.

Haladyna, T. M., & Downing, S.M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*. 23(1), 17-27.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research.* 77(1): 81-112.

Herbers, J., Noel, G., Cooper, G., Harvey, J., Pangaro, L., & Weaver, M. (1989). How accurate are faculty evaluations of clinical competence? *Journal of general Internal Medicine.* 4(3), 202-208.

Hodges, B. (2003). Validity and the OSCE. *Medical Teacher*. 25(3): 250-254.

Hodges, B. (2013). Assessment in the post-psychometric era: Learning to love the subjective and collective." *Medical Teacher.* 35(7), 564-568

Hodges, B., Lingard, L. (2012). The Question of Competence: Reconsidering Medical Education in the Twenty-First Century. Ithaca, New York, Cornell University Press.

Hodges, B, Regehr, G., McNaughton, N., Tiberius, R., & Hanson, M. OSCE checklists do not capture increasing levels of expertise. *Academic Medicine.* 1999. 74(10), 1129-1134.

Holmboe, E. S. (2004). Faculty and the observation of trainees' clinical skills: problems and opportunities. *Academic Medicine.* 79(1), 16-22

Huwendiek, S., Mennin, S., Dern, P., Friedman Ben-David, M., Van Der Fleuten, C., Tonshoff, B., Nikendei, C. (2010). Expertise, needs and challanges of medical educators: Results of an international web survey. *Medical Teacher*. 32(11), 912-918.

Kahneman, D. (1973). Attention and effort. Attention and Consciousness in Pyschology, in *Philosophy and Cognitive Science*. Englewood Cliffs, NJ: Prentice Hall.

Kane, M. (2006). Validity Educational Measurement. R. L. Brennan. Westport, CT: Praeger Pub Text.

Kane, M. (1992). The assessment of professional competence. *Evaluation & the health professions*. 15(2), 163-182.

Kim, J., Neilipovitz, D., Cardinal, P., Chiu, M., & Clinch, J., (2006). A pilot study using high-fidelity simulation to formally evaluate performance in the resuscitation of critically ill patients: The University of Ottawa Critical Care Medicine, High-Fidelity Simulation, and Crisis Resource Management I Study. *Critical Care Medicine*. 34(8): 2167-2174.

Kogan, J., & Holmboe, E. (2013). Realizing the Promise and Importance of Performance-Based Assessment. *Teaching and Learning in Medicine*. 25(1), 68-74.

Kogan, J., Conforti, L., Bernabeo, E., Iobst, W. & Holmboe, E. (2011). Opening the black box of clinical skills assessment via observation: a conceptual model. *Medical Education*. 45(10): 1048-1060.

Kogan, J., Hess,B., Conforti, L., & Holmboe, E. (2010). What Drives Faculty Ratings of Residents' Clinical Skills? The Impact of Faculty's Own Clinical Skills. *Academic Medicine*. 85(10), 25-32.

Kohn, L., Corrigan, J.M., & Donaldson, M.S. (2000). To err is human: building a safer health system. A report of the Committee on Quality of Health Care in America, *Institute of Medicine*. Washington, DC: National Academy Press.

Kuper, A. & D'Eon, M. (2011). Rethinking the basis of medical knowledge. *Medical Education*. 45(1), 36-43.

LaMantia, J., Rennie, D., Risucci, R., Cydulka, L., Spillane, L., Graff, J., Becher, k., & Kleinschmidt, K. (1999). Interobserver Variability among Faculty in Evaluations of Residents Clinical Skills. *Academic Emergency Medicine*. 6(1), 38-44.

Landy, F.J., Farr, J.L. Performance rating. *Psychological Bulletin*. (1980). 87(1), 72-107.

Lavie, N. (1995). Perceptual load as a necessary condition for selective attention. *Journal of Experimental Psychology: Human Perception and Performance*. 21(3), 451-468.

Leape, L.L., Fromson, J.A., Problem doctors: is there a system-level solution? *Annals of Internal Medicine*, 2006. 144(2), 107-115.

Lurie, S. J., Mooney, C.J., & Lyness, J.M. (2011). Commentary: Pitfalls in Assessment of Competency-Based Educational Objectives. *Academic Medicine.* 86(4): 412-418.

Lurie, S., Mooney, C., Lyness, J. (2009). Measurement of the general competencies of the Accreditation Council for Graduate Medical Education: a systematic review. *Academic Medicine*, 84(3), 301-309.

Margolis, M., Clauser, B., Cuddy, M., Ciccone, A., Mee, J., Harik, P., & Hawkins, R. (2006). Use of the mini-clinical evaluation exercise to rate examinee performance on a multiple-station clinical skills examination: a validity study. *Academic Medicine.* 81(10): S56-S60.

Mayer, R. E. (1999). The promise of educational psychology. Vol. 2, *Learning in the Content Areas*. Upper Saddle River, NJ: Prentice Hall.

Mayer, R. E., Chandler, P. (2001). When learning is just a click away: Does simple user interaction foster deeper understanding of multimedia messages? *Journal of educational psychology*. 93(2): 390-397.

Mayer, R. E., Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist.* 38(1): 43-52.

Mayer, R. E., Heiser, J., & Lonn, S. (2001). Cognitive constraints on multimedia learning: When presenting more material results in less understanding. *Journal of Educational Psychology*. 93(1): 187-198.

Melchers, K. G., Kleinmann, M., & Prinz, M.A. (2010). Do Assessors Have Too Much on their Plates? The Effects of Simultaneously Rating Multiple Assessment Center Candidates on Rating Quality. *International Journal of Selection and Assessment* 18(3): 329-341.

Miller, G. (1990). The assessment of clinical skills/competence/performance. *Academic Medicine*. 65(9): S63-S67.

Morgeson, F., Campion, M. (1997). Social and cognitive sources of potential inaccuracy in job analysis. *Journal of Applied Psychology*. 82(5): 627-654.

Murphy, K, Philbin, A., & Adams, S.R. (1989). Effect of Purpose of Observation on Accuracy of Immediate and Delayed Performamce Ratings. *Organizational Behaviour and Human Decision Processes.* 43(3): 336-354.

Murphy, K., Cleveland, J., Skattebo, A., & Kinney, T. (2004). "Raters who pursue different goals give different ratings." *Journal of Applied Psychology*. 89(1): 158-164.

Newlin-Canzone, E., Scerbo, M.W., Gliva-McConvey, G., & Wallace, A.M. (2013). The Cognitive Demands of Standardized Patients: Understanding Limitations in Attention and Working Memory With the Decoding of Nonverbal Behaviour During Improvisations. *Simulation in Healthcare.* 8(4): 207-214.

Nisbett, R., Wilson, T. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review* 84(3): 231-29.

Noel, G., J. Herbers, M. Caplow, G. Cooper, L. Pangaro and J. Harvey (1992). How well do internal medicine faculty members evaluate the clinical skills of residents? *Annals of Internal Medicine.* 117(9): 757-765.

Norcini, J. J. (2005). Current perspectives in assessment: the assessment of performance at work. *Medical Education.* 39(9): 880-889.

Norcini, J., Blank, L., Arnold, G., & Kimball, H. (1995). The mini-CEX (clinical evaluation exercise): A preliminary investigation. *Annals of Internal Medicine.* 123(10): 795-799.

Norman, G., van der Vleuten, C.P. & Graaff, E. (1991). Pitfalls in the pursuit of objectivity: issues of validity, efficiency and acceptability. *Medical Education.* 25(2): 119-126.

Paas, F. & Van Merriënboer, G. (1994). Measurement of cognitive load in instructional research. *Perceptual and Motor Skills*.79(1): 419-430.

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist*. 38(1): 1-4.

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

Paas, F., van Gog, T., & Sweller, J. (2010). Cognitive load theory: New conceptualizations, specifications, and integrated research perspectives. *Educational Psychology Review*. 22(2): 115-121.

Plass, J. L., Moreno, R., & Brünken, R. (2010). Cognitive load theory. New York, NY: Cambridge Univ Pr.

Regehr, G., MacRae, H., Reznick, R., & Szalay, D. (1998). Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Academic Medicine*. 73(9): 993-997.

Roch, S. G., Woehr, D,J., Mishra, V., & Kieszczynska, U. (2011). Rater training revisited: An updated meta-analytic review of frame-of-reference training. *Journal of Occupational and Organizational Psychology*. 85(2): 370-395.

Roodenrys, K., Agostinho, S., Roodenrys, S., & Chandler, P. (2012). Managing one's own cognitive load when evidence of split attention is present. *Applied Cognitive Psychology*. 26(6): 878-886.

Rushton, A. (2005). Formative assessment: a key to deep learning? *Medical Teacher*. 27(6): 509-513.

Saedon, H., Salleh, S., Balakrishnan, A., Imray, C.H., & Saedon, M. (2012). The role of feedback in improving the effectiveness of workplace based assessments: a systematic review. *Medical Education.* 12(1): 25.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*. 78(1): 153-189.

Simon, H. A. (1972). Theories of bounded rationality. In *Decision and Organization*. (pp.161-176). Amsterdam: North Holland.

Streiner, D., Norman, G. (2008). Health measurement scales: a practical guide to their development and use. New York, NY: Oxford University Press.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*. 12(2): 257-285.

Tavares, W., & Eva, K.W. (in progress). The Impact of Rating Demands on Rater Based Assessment of Clinical Competence.

Tavares, W. & K. W. Eva (2012). Exploring the impact of mental workload on rater-based assessments. *Advances in Health Sciences Education*. 18(2): 291-303.

Tavares, W., Ginsburg, S., & Eva, K.W. (2014). (submitted) Selecting and Simplyfying: Rater Behaviours and Performance when Considering Multiple Competencies.

Tavares, W., LeBlanc, V.R., Mausz, J., Sun, V., & Eva, K.W. (2013). Simulation Based Assessment of Paramedics and Performance in Real Clinical Contexts. *Prehospital Emergency Care*.17(1): 57-67.

Tavares, W., S. Boet, R. Theriault, T. Mallette., & Eva, K.W. (2012). Global Rating Scale for the Assessment of Paramedic Clinical Competence. *Prehospital Emergency Care* 17(1): 57-67.

Tsang, P. S., & Vidulich, M.A. (2006). Mental workload and situation awareness. Hoboken, NJ, Wiley and Sons.

Van Der Vleuten, C.P., & Schuwirth, L.W. (2005). Assessing professional competence: from methods to programmes. *Medical Education.* 39(3): 309-317.

Van der Vleuten, C., L. Schuwirth, F. Scheele, E. Driessen & B. Hodges (2010). The assessment of professional competence: building blocks for theory development. *Best Practice & Research Clinical Obstetrics & Gynaecology.* 24(6): 703-719.

van der Vleuten, C. (1996). The assessment of professional competence: developments, research and practical implications. *Advances in Health Sciences Education* 1(1): 41-67.

van der Vleuten, C., Norman, G., & Graaff, E. (1991). Pitfalls in the pursuit of objectivity: issues of reliability. *Medical Education*. 25(2): 110-118.

Van Merriënboer, J., & Sweller, J. (2010). Cognitive load theory in health professional education: design principles and strategies. *Medical Education* 44(1): 85-93.

Van Merriënboer, J., Kirschner, P., & Kester, L. (2003). Taking the load off a learner's mind: Instructional design for complex learning. *Educational Psychologist.* 38(1): 5-13.

Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical issues in ergonomics science.* 3(2): 159-177.

Wickens, C. D. (2008). Multiple resources and mental workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society.* 50(3): 449-455.

Wickens, C. D., Carswell, C. (2006). Handbook of human factors and ergonomics, Third Edition. Hoboken, NJ, Wiley

Williams, R., Klamen, D., & McGaghie, W. (2003). Cognitive, Social and Environmental Sources of Bias in Clinical Performance Ratings. *Teaching and Learning in Medicine.* 15(4): 270-292.

Wind, L., Van Dalen, J., Muijtjens, A., & Rethans, J. (2004). Assessing simulated patients in an educational setting: the MaSP (Maastricht Assessment of Simulated Patients). *Medical Education.* 38(1): 39-44.

Woehr, D., & Huffcutt, A. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology.* 67(3): 189-205.

Wood, T. J. (2013). Exploring the role of first impressions in rater-based assessments. *Advances in Health Sciences Education*: ePublished Mar 26. [doi: 10.1007/s10459-013-9453-9]

Yeates, P., O'Neil, P., Mann, K., Eva, K.W. (2012). Effect of Exposure to Good vs Poor Medical Trainee Performance on Attending Physician Ratings of Subsequent PerformancesAttending Physician Ratings of Performance. *JAMA.* 308(21): p. 2226-2232.

Yeates, P., O'Neill, P., Mann, K., Eva, K. (2012). Seeing the same thing differently. *Advances in Health Sciences Education,* 18(3), 325-341.

Yeates, P., P. O'Neill, K. Mann and K. W Eva (2013). You're certainly relatively competent': assessor bias due to recent experiences. *Medical Education.* 47(9): 910-922.

Yurko, Y., Scerbo, M., Prabhu, A., Acker, C., Stefanidis, D. (2010). "Higher Mental Workload is Associated With Poorer Laparoscopic Performance as Measured by the NASA-TLX Tool." Simulation in Healthcare 5(5): 267-271.