uOttawa

L'Université canadienne
Canada's university

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTOCTORALES

uOttawa

L'Université canadienne
Canada's university

FACULTY OF GRADUATE AND POSDOCTORAL STUDIES

Sydney Smee
AUTEUR DE LA THÈSE / AUTHOR OF THESIS

Ph.D. (Education)
GRADE / DEGREE

Faculty of Education
FACULTÉ, ÉCOLE, DÉPARTEMENT / FACULTY, SCHOOL, DEPARTMENT

Comparing Scoring Instruments for the Performance of Professional Competencies

TITRE DE LA THÈSE / TITLE OF THESIS

Dany Laveault
DIRECTEUR (DIRECTRICE) DE LA THÈSE / THESIS SUPERVISOR

CO-DIRECTEUR (CO-DIRECTRICE) DE LA THÈSE / THESIS CO-SUPERVISOR

EXAMINATEURS (EXAMINATRICES) DE LA THÈSE / THESIS EXAMINERS

B. Cousins

D. Trumpower

K. Eva

M. Simon

Gary W. Slater
Le Doyen de la Faculté des études supérieures et postdoctorales / Dean of the Faculty of Graduate and Postdoctoral Studies

Comparing Scoring Instruments for the Performance Assessment of Professional Competencies

Sydney M. Smee

Thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
In partial fulfillment of the requirements
For the PhD in Education (Measurement and Evaluation)

Educational Studies Program
Faculty of Education
University of Ottawa

Your file Votre référence
ISBN: 978-0-494-34149-0
Our file Notre référence
ISBN: 978-0-494-34149-0

Table of Contents

List of Tables

List of Figures

Legend

Acronym	Definition
CL-1	Checklist - Depression
CL-2	Checklist - Delirium
CRS-1	Case-specific Rating Scale - Depression
CRS-2	Case-specific Rating Scale - Delirium
OSCE	Objective Structured Clinical Examination
PIRS	Patient Interaction Rating Scale
SRS-1	Skill-specific Rating Scale - Depression
SRS-2	Skill-specific Rating Scale - Delirium

Abstract

Performance assessments of professionals are commonly scored with rating scales but checklists are used with the Objective Structured Clinical Examination (OSCE). Theory suggests checklists are too detailed (Norman, 2005a) and studies show that generic rating scales (e.g., Hodges, Regehr, McNaughton, Tiberius, & Hanson, 1999) are more discriminating and reliable. However, generic scales may represent clinical expertise too narrowly, resulting in less valid scores.

Study One asked how *case-specific* rating scale scores and decisions compared to checklist scores and decisions at representing professional competency. Study Two asked the same question about a *skill-specific* rating scale.

Data were from a medical licensure OSCE. Participants were 1,587 test takers and 190 physician raters. Two patient cases (Depression and Delirium) were used. In each case, two physicians scored. One scored a checklist and the other a rating scale; both scored a patient interaction rating scale (PIRS).

Results from both studies showed internal consistency was higher for rating scales (e.g., for Study One α=.87 and α=.79 for rating scales; α=.55and α=.64 for checklists). Item-total score correlations (ITC) for each rating scale as an item within the OSCE were also higher than the ITC for the checklists in both studies. Logistic regression analyses predicting pass/fail decisions from a 3-variable expertise model explained more variance for the rating scales decisions than it did for the checklists in both studies (e.g., R_L^2=16.4% and R_L^2=16.4% for rating scales; R_L^2=12.7% and R_L^2=5.7% for checklists in Study One). The highest Pearson correlations (corrected for attenuation) were between the rating scale scores and the respective PIRS scores.

In conclusion, these rating scale scores are more discriminating and reliable than checklist scores but correlations with PIRS scores indicate they may not measure the intended dimension of the clinical expertise construct. Evidence for a validity argument was strongest for the case-specific rating scale for Depression, raising the question of whether rating scale methodology is appropriate for all OSCE cases.

Acknowledgments

I began this journey because David Blackmore, Ph.D. said I would and he never let me think otherwise. He has my heartfelt appreciation for his belief in me and the mentoring he gave so generously along the way.

I am also indebted to Dale Dauphinee, M.D., FRCPC because he made my excuses for not starting disappear.

Many thanks are owed to my advisor, Prof. Dany Laveault and my thesis committee, Drs. Bradley Cousins, Carol Miles and Marielle Simon for always insisting on greater clarity and for their patience during the slow course of this journey.

Friends, family and colleagues have supported me all along the way. I know they are grateful that this part of my journey is drawing to a close and I thank them all for seeing me to its end. For their timely advice at critical times, I am grateful to Tom Maguire, Ph.D. and Andrija Popovic, M.A. Special thanks go to my sister Sue for her ability to critically edit without drawing blood and because she chanted for me.

This research could not have been completed without the permission and the funding provided by the Medical Council of Canada and I am very thankful to the Council for this generous support.

Introduction

This research is about high stakes performance assessment of professional competency; more specifically, it compares the validity of checklist scores to that of rating scale scores and assesses which scoring instruments lead to the most defensible pass-fail decisions. Because the stakes are high, performance assessments of professionals such as doctors and pilots are expected to meet rigorous psychometric standards of validity and reliability.

Psychometric theory evolved with and supported the widespread use of standardized tests in North American education. Negative consequences attributed to standardized tests have emerged, however, stimulating a growing interest in the use of performance assessments such as those focused on in this research. Performance assessments do not meet traditional psychometric standards and some would argue, should not. Nonetheless, in some environments, meaningful compromises are being found that reconcile traditional measurement criteria and performance assessment. This research is about seeking the best compromise in one such environment, high stakes performance assessment for medical licensure in Canada but the expectation is that progress made in this domain may benefit performance assessments in other domains. The results contribute to the level of confidence that test users and test takers can have in the results of the instruments used to score high stakes performance assessment.

While rating scales and rubrics are used for high stakes performance assessment of professional competency in several fields, checklists are more common in the health professions. This reliance on checklists can be attributed to the extensive use of the

Objective Structured Clinical Examination (OSCE). The purpose of the OSCE is to objectively measure clinical skills such as history taking, physical diagnosis, doctor-patient communication, and patient management based on test taker performance as demonstrated over a series of patient simulations. The first OSCE was run over 30 years ago (Harden, Stevenson, Downie, & Wilson, 1975) and it introduced the use of standardized scoring to the assessment of clinical skills. The reliability and validity of the checklist scores contributed greatly to the OSCE gaining acceptance (van der Vleuten & Swanson, 1990).

At the time of that first OSCE, clinical expertise was conceptualized as a problem-solving construct based on skills that transferred across patient problems. This construct has since given way to a clinical reasoning construct. The latter is based on the knowledge, both formal and experiential, that physicians must acquire if they are to make sound judgments about patient care. Within the clinical reasoning construct, expertise develops if a physician's formal knowledge base is expanded by extensive experience with a range of patient problems (Norman, 2005b). OSCE checklist scores fit well with a skill-based construct but more recently have been challenged as being too detailed for a clinical reasoning construct. Empirical studies support this challenge. As a result, rating scales have been introduced to the OSCE as adjuncts to checklists for scoring limited aspects of clinical expertise.

OSCE-based studies that compared rating scale scores to checklist scores found rating scales scores to be superior because they more consistently discriminated between levels of training. However, these studies faced some significant methodological limitations, as well as focusing solely on generic rating scales. A generic instrument is

highly appealing because it greatly reduces development costs but leads to the concern that motivated this research. Although generic rating scales have considerable utility, they may not capture some critical aspects of expertise as defined by a clinical reasoning construct, leading to a loss of validity. This research entails two studies that compare scores and pass-fail decisions from specific rating scales (i.e., specific to a patient case or to a clinical skill) to checklist scores. In effect, the utility of a generic rating scale is relinquished for the potential gain in construct validity that a case- or skill-specific rating scale may offer.

The Literature Review chapter describes the historical, theoretical and empirical setting for two parallel studies: Study One, which is based on case-specific rating scales and Study Two, which is based on a skill-specific rating scale. The research questions are presented, along with the study design. The Methods chapter provides details regarding the participants, the development of the study instruments, the data collection process, and the data analyses. The third chapter presents the results in three parts. The first part focuses on item analyses that are organized by instrument and grouped by study. The second part compares the validity of the pass-fail decisions from different instruments using a model of expertise and regression analyses. These results are organized by study. The third part is a summary of the feedback received from the physicians who scored the study instruments. The Discussion chapter examines the results relative to the findings of previous studies, considers their limitations and makes further recommendations on the best possible use of checklists and rating scales.

Literature Review

The literature review is in five parts, each one contributing to the framework for the research questions. The first part provides an historical context that explains the theory and ideals underlying performance assessments in the field of education today. The second part discusses the psychometric criteria of validity and reliability that determine the merit of an educational assessment. The focus of this part is on those aspects of the criteria that are most germane to scoring high stakes performance assessment. The third part looks at current examples of high stakes performance assessment to illustrate how reliable scores are attained. The fourth part is a discussion of clinical expertise theory, which defines the relevant domain for this particular research. Fifth and finally, is a critique of other studies that have compared checklist scores to rating scale scores within the clinical domain.

The framework is drawn in part from the literature review. It also incorporates my concern that some critical and specific aspects of performance that are captured by checklist scores may be lost or poorly represented by the generic rating scales used in other studies. This concern has arisen from discussions I have had with many of the test takers who fail a performance-based medical licensure assessment. A rating scale may be the most reliable instrument for scoring performance assessments, but what the items within a rating scale should represent requires further investigation.

An Historical Perspective on Educational Testing

For most of the twentieth century educational testing in North America was based on principles drawn from ideas that emerged early in the century. Three of the most

significant influences in education were the social efficiency movement, behavioural learning theory and scientific measurement (Shepard, 2000). For instance, when social efficiency principles that had been developed to improve factory productivity were applied to education, separate curriculum tracks for students of different abilities were implemented. Streaming students by their assumed ability level was a utilitarian strategy aimed at efficiently educating each student according to his innate abilities. The focus was on providing productive citizens to an increasingly industrial society, not on the learning needs of the students.

Behavioural learning theory defined learning as a process of accumulating knowledge in a linear, step-by-step manner. Within the behavioural model, teachers were responsible for transferring knowledge in a sequenced approach and for testing frequently to ensure that sufficient learning had occurred before they allowed learners to move on to the next step. The behavioural model of learning was influential into the 1970s (Haertel, 1999).

By the end of the 1920s, compulsory school attendance laws were in effect and greatly expanded the number of students in the educational system. The confluence of behavioural theory, social efficiency values and the pressures created by the rapidly expanding school system prompted the introduction of the objective paper-and-pencil test (Stiggins, 1991). The use of objective tests, largely based on multiple-choice questions, marked a distinctive shift away from traditional, teacher-centred testing toward a standardized approach that relied on scientific principles and centralized test development. The use of objective testing was reinforced by the emerging field of

psychological measurement which sought to apply scientific methods to testing intelligence and educational achievement (Stiggins, 1991).

Objective tests were easily mass produced and with the introduction of optical scanning technology, easily scored (Eisner, 1999). They offered a fair way of testing large numbers of test takers and produced scores that seemed ideal for comparing individuals and schools. In Shepard's (2000) words, "dominance of the 'objective test' ... has been the single most striking feature of achievement testing in the United States from the beginning of the century to the present day" (p. 5). The wealth of measurement research and technical sophistication that came with objective testing further widened the gap between teachers and assessment because testing more and more became the purview of measurement experts (Stiggins, 1991).

The entrenchment of objective testing in the American educational system made large scale testing standard practice (Stiggins, 1991); public and media concerns about the quality of education made large scale accountability testing essential in the eyes of policy makers. Objective tests with standardized administration were an affordable, easily mandated means of generating reports to satisfy the demand from the public and the policy makers that schools and teachers be held accountable for the education of the students entrusted to their care (Linn, 2000). As the use of standardized testing grew in the United States, it also spread to Canadian jurisdictions (McEwen, 1995; Wilson, 1999).

By the 1980s standardized tests dominated North American educational measurement, a trend that continued into the 1990s with results that commonly showed upward trends in student achievement. The problem, as described by Linn (2000) and

others, was that scores from these tests gave an inflated view of student achievement. The score inflation was attributed to poor testing practices, such as excluding poor students from the testing, and to teaching that focused too narrowly on what was being tested, thereby promoting rote learning. While standardized testing marked a significant step forward in fairly assessing large numbers of students, the pressure on teachers and schools to produce students that performed well on standardized tests also had negative consequences. As understanding of the negative consequences grew, more attention was paid to the shortcomings of standardized testing (Wilson, 1999).

A societal assumption is that learning should lead to success beyond the classroom. Therefore, the ideal assessment should indicate how well learners will perform in the larger world, especially the labour market, and not just show how well they performed in school. Scores from standardized tests did not meet this ideal; for example, they proved to be poor predictors of job performance, which is a significant measure of success outside the classroom (Sternberg, 1998). Moreover, standardized test scores were only weakly related to the quality of schooling because so many other factors contribute to student achievement; factors like genetic predisposition, home life, and exposure to language and to television (Stake, 1999).

Wiggins (1989) was an early and influential critic of large-scale standardized testing. He called for authentic assessments that would be central to learning; for assessments that would honour the role of human judgment and dialogue in determining student achievement. Such assessments, he argued, would be based on tasks that reflected the challenges found in the workplace and in daily living. He contended that one test score, based on a series of decontextualized multiple choice questions, could not possibly

be a valid representation of student achievement. He was not alone. Others, such as Stiggins (1994) and Shepard (2000) also argued for classroom-based assessment strategies that would foster a learning culture.

Standardized testing was principally based on behavioural learning theory. As dissatisfaction with standardized testing grew, so did a shift away from behavioural thinking. Interest in cognitive learning theory began to flourish and is the underpinning for the authentic or alternative, performance-based approaches to assessment now being promoted. Cognitive theory defines learning as the construction of meaning that occurs as the learner interprets and clarifies his experience of the world (e.g., Delandshere & Petrosky, 1994; Supovitz & Brennan, 1997). From the cognitive perspective, the acquisition of new knowledge is a process of structuring and linking new information with existing knowledge structures (Snow & Lohman, 1993) and the context in which learning occurs plays a significant role in shaping what is learned, as do the learner's prior knowledge and beliefs. Authentic assessment, based on whole tasks lodged in a context relevant to the learner and integrated into the classroom experience, is an expression of cognitive learning principles. Authentic assessments are believed to engage higher order thinking processes and to draw on multiple skills; and therefore, they are deemed to be more valid measures of learning (Resnick & Resnick, 1992).

With a performance-based approach to assessment a multitude of formats, such as the completion of tasks, participation in group exercises, oral presentations, and the compilation of portfolios, become acceptable means of measuring learning. The more closely the assessment tasks are connected to workplace activities and life tasks, the more authentic the performance assessment. However, authentic assessment as conceptualized

by Wiggins (1989) is not a testing strategy that can be tacked on at the end of a course. In fact, he argues that schooling should be restructured to incorporate assessment strategies rooted in the performance of exemplar tasks that represent not just the curriculum but also the ambiguity and context of challenges that learners face outside the classroom. While many performance assessments do not fully meet this standard of authenticity, the use of test questions and tasks that invoke more than recall is increasingly widespread.

The move to more authentic, performance-based assessments comes with a challenge. Although more authentic or direct forms of assessment have the potential for higher validity, there needs to be evidence to support this claim, especially when such assessments are used for making high stakes decisions. The standards of validity and reliability developed over decades of standardized testing are as much value statements about fair testing as they are technical requirements. Understanding the nature of validity and reliability is important to understanding the criteria used to compare scoring instruments and so the aspects of validity and reliability most germane to this research are discussed next.

Psychometric Criteria: Validity and Reliability

Performance assessments overcome some of the limitations of standardized tests by using complex tasks reflective of real world challenges rather than relying on selected answer formats like the multiple choice question, but they also introduce complications. Generating reliable scores is one of them. Unlike selected answer formats, which are usually machine scored, human observation and interpretation are required for scoring performance assessments (Haertel, 1999; Madaus & O'Dwyer, 1999). For example, scoring the level of skill demonstrated in the performance of a task or determining the

quality of a portfolio relies upon human judgment, which is summarized by completing a scoring instrument. Scoring instruments are designed according to a defined construct and the items within an instrument inform the rater of the construct that they are expected to assess.

Defining the construct that is being measured is the first step toward creating a valid assessment and is fundamental to developing meaningful scoring instruments. Classically, a construct is a theory about the nature of a latent trait and its relationship to observable variables; for example, the relationship between intelligence and school performance on tasks of verbal or quantitative reasoning (Bracey, 2005). Although such trait-based constructs are useful in many areas of research, cognitive theory challenges the assumption that latent traits or abilities sufficiently explain educational achievement and professional competencies (Ericsson & Charness, 1994; Haertel, 1999; Mislevy, 1993; Snow & Lohman, 1993; Sternberg, 1998; van der Vleuten, 1996). As a result, educational constructs are developed, more appropriately, from theories about the cognitive processes that are required in a field such as mathematics (Nichols & Sugrue, 1999) or from theories about the roles that define a profession such as medicine (Schuwirth & van der Vleuten, 2004).

If a construct is not well articulated, then an assessment may under-represent the construct (i.e., too narrow an understanding) or the assessment may capture construct-irrelevant variance (i.e., too broad an understanding). Either effect compromises the validity of the intended inferences (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). Nichols and Sugrue (1999) argue that a critical concern is the tendency for scoring

instruments to under-represent the rich complexity of ability demonstrated by test takers during a performance assessment. They are particularly troubled by rubrics that lead to a single score. Such rubrics may oversimplify the construct and force raters to fit their judgments into a few abstract categories that minimize significant differences in observed performance. The risk with under-representing a complex construct is that the loss of content relevance will be sufficient to undermine the validity of the scores and therefore of the inferences drawn from them. Scoring instruments for summative assessments are particularly vulnerable to under-representing complex constructs because reliable scores are essential and are most easily achieved with unidimensional instruments that lead to a single score for each test taker.

While reliable scores are essential to high stakes summative assessments, there are serious consequences if reliability is accomplished by sacrificing validity. This is especially true for certification and licensure decisions where the consequences affect not only the individual professional, but also their profession as a whole and the public who rely on their services. Approaches to balancing validity and reliability criteria for performance-based certification and licensure assessments can be found in a number of fields; such as teaching (Committee on Assessment and Teacher Quality, 2001b), flying (Baker & Dismukes, 2002), law enforcement (Centrex, 2006b), and the health professions (e.g., Austin, O'Byrne, Pugsley, & Quero, 2003; Parker-Taillon, Cornwall, Cohen, & Rothman, 1992; Reznick et al., 1993). The high cost and significant effort required to develop reliable assessments in these fields is justified with validity arguments. Inferences about workplace performance drawn from assessments constructed of job-related tasks are more defensible than inferences made from paper-and-pencil type

tests. Although this argument is intuitively sound, the inferences are only defensible if they lead to pass-fail decisions that consistently and accurately reflect the observed performance and lead to higher predictive validity of job performance or success within a profession.

Miller (1990) describes four levels at which assessment can occur. The *knows* level is first and largely refers to factual information. Next is the *knows how* level, which refers to the ability to solve problems and describe procedures, then there is the *shows how* level that requires a demonstration of skills. Lastly, there is the *does* level, referring to what is actually done in the workplace. Performance assessments measure competencies at the *shows how* level. This means inferences are made from *shows how* to *does,* a shorter distance than the one from the *knows* level to the *does* level. A shorter distance means a narrower inferential gap and provides support for a validity argument.

Figure 1 illustrates the differences in distances between the test, the score and the inferences for standardized tests and performance assessments. With the multiple-choice questions common to standardized tests, there is no gap, meaning no interpretation and almost no confusion, between performance on the item (i.e., selecting an option from a list of answers) and the score, as shown by the nested circles in the upper left of Figure 1. There is however a large gap between the score and any inferences that are drawn regarding the test taker's performance in a specific domain of competency, shown by the line travelling across the upper portion of Figure 1. This gap is created by scoring error related to test takers guessing or to poorly constructed items, and by the conceptual distance that lies between the *knows* level and the *does* level.

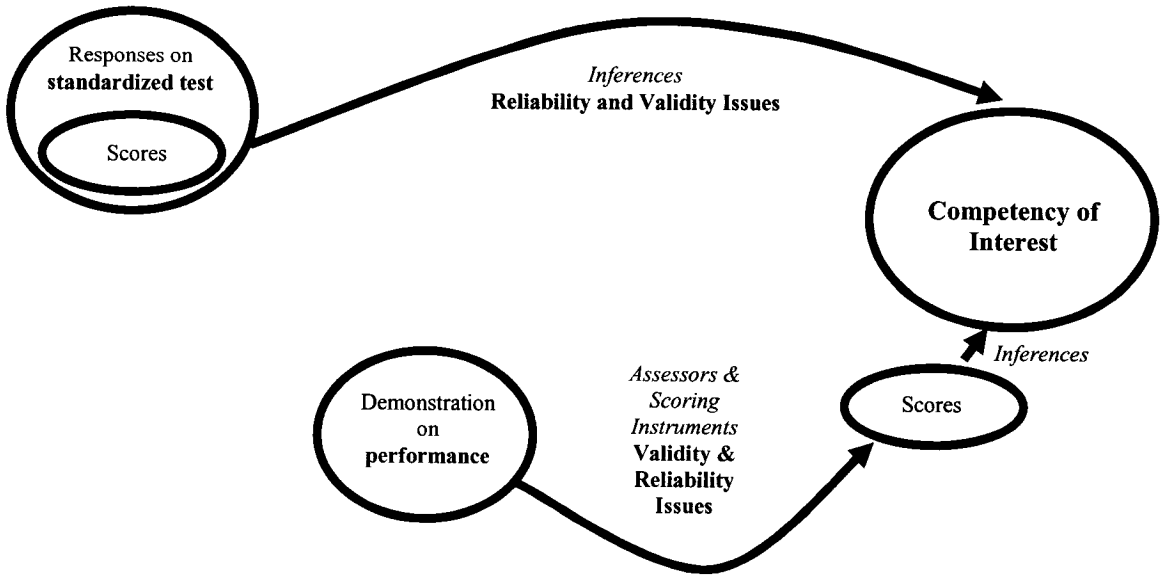


Figure 1 Comparison of standardized tests and performance assessments

With performance assessment the gap between performance on the task, the *shows how* level, and the competency of interest, the *does* level, is smaller, as shown in the lower portion of the figure. However, there is a gap between performance and scores that is created by the error introduced when an assessor summarizes a complex performance with a single score or with a few points on a rubric. This gap is represented by the line traveling from performance in the lower left portion of the figure to scores on the lower right. The smaller distance from scores to inferences is the basis for the claim to authenticity.

By using tasks that elict the kind of performance that is expected in the competency of interest, performance assessments increase the *potential* for validity. Messick (1994) interprets this as a validity argument based on the belief that standardized tests under-represent educational constructs because they decontextualize and decompose achievement into components. The claim is that performance assessments better represent

the complexity of achievement and therefore minimize construct under-representation. Scores are assumed to be more valid because they are believed to better represent construct-relevant variance (Messick, 1994; Nichols & Sugrue, 1999; Supovitz & Brennan, 1997). Gitomer (1993) emphasizes that the link between assessment *tasks* and the world outside the classroom is strengthened when the tasks or problems are presented within a context that simulates real world problems where resources and timelines constrain the range of possible solutions. The more the nature of an assessment reflects real world tasks, the more direct and authentic the performance assessment. Directness minimizes the degree of interpretation required to make inferences from observed performance to real world tasks. Messick (1994) defines this directness as a validity argument based on minimizing construct-irrelevant variance.

However, validity is ultimately determined by the degree to which the *scores* represent the observed performance. Authentic tasks are not sufficient to establish the validity of an assessment. If an assessment is to be more than just a showcase of student work, it must be summarized in a meaningful way, with minimal loss of information. To do less is to limit construct representation. The goal is "scoring rubrics that are neither specific to the task nor generic to the construct but are in some middle ground reflective of the classes of tasks that the construct empirically generalizes or transfers to" (Messick, 1994, p. 17).

The development of valid scoring instruments for educational assessment is undermined when a latent ability is assumed to underlie performance because the assumption leads to unidimensional instruments being used to assess what are actually multidimensional constructs (Nichols & Sugrue, 1999). Cognitive theorists have argued

that educational constructs are multidimensional combinations of processes, concepts and procedures. In response, educators have developed performance assessments to sample these constructs. However, if the interaction of the dimensions within such a construct is not taken into account by the scoring, then test takers who complete a task via different reasoning and procedural approaches may perform quite differently but end up with the same score. As Nichols and Sugrue have argued, a single score from a unidimensional instrument cannot describe such differences, even if it is a reliable score.

Although unidimensional scales leading to a single score generally produce reliable scores, using single scores to represent complex, multidimensional performances creates the real risk of creating a composite score that is less reliable than separate scores for each dimension, especially when such dimensions show limited correlations among themselves (Nichols & Sugrue, 1999). Even though the tasks may allow the rater to reliably discriminate between test takers who use different concepts and reasoning strategies to complete the task, much of this information is lost when the rater's judgments are reduced to a single score. Loss of true score variance leads to less discrimination and lower estimates of reliability.

At the same time, the influence of the traditional measurement assumption that scores must be reliable or consistent encourages the use of unidimensional scoring instruments because procedures for estimating the reliability of these scores are well-established. Delandshere and Petrosky (1998) challenge the measurement assumption that scores must be quantitatively reliable by arguing that scoring instruments that meet this standard limit the validity of educational assessments. Accepting the assumption that quantitatively reliable scores are more defensible than scores that meet alternative,

qualitative concepts of reliability ignores the negative impact of a quantitative scoring process on the elaborate interpretative process of a rater, making it both implicit and invisible.

A further argument for moving away from established measurement standards is the tendency for conventional reliability analyses to treat the interaction between item and test takers as relative error (Nichols & Smith, 1998; Nichols & Sugrue, 1999). Treating item-test taker interaction as relative error ignores cognitive explanations of learning which suggest that this interaction may be construct-relevant variance that reflects differences in test taker approaches to different tasks.

While the debate over how best to determine reliability when assessing competencies from complex constructs may eventually result in more qualitative and valid approaches to summarizing performance assessments, current standards of fairness mean that traditionally reliable scores are still critical for making defensible, high stakes pass-fail decisions (American Educational Research Association et al., 1999). Moreover, in the context of high stakes assessment, there are times when a certain degree of authenticity may be traded for higher reliability (Brennan & Johnson, 1995). What follows are examples of performance assessments of professional competence that illustrate the kind of trade-offs being made to achieve score reliability when the stakes are high.

High Stakes Performance Assessment

High stakes assessments are those that lead to significant decisions about individuals, teachers, schools, or policy. In this research, the focus is on performance assessment as used to make decisions about individuals. Licensure and certification

assessments are high stakes assessments of an individual's professional competence. Licensure requirements are imposed by governments to protect the public so that only those individuals who can demonstrate sufficient knowledge, skill and judgment to practice safely and effectively in a given field are allowed to do so (American Educational Research Association et al., 1999). Licensure assessments commonly occur at the end of training programs. In some professions, such as law and medicine, this means large numbers of individuals have to be assessed in a very short time period. The time constraint is one of the logistical challenges of licensure testing. Certification is a voluntary process verifying that individuals have met a certain level of knowledge and skill in a profession (American Educational Research Association et al., 1999), although certification may also be required for professional advancement. Assessment for certification can occur at different times in an individual's career path, allowing testing bodies to administer the assessments on an ongoing basis.

The level of competency being assessed for certification is often at the advanced level of professional achievement required for promotion whereas licensure is assessing the level of competency required for entry into practice. That being said, certification and licensure assessments have much in common, most especially their outcomes. Each leads to a dichotomous decision to certify or not, to be eligible for licensure or not; in plain language, to pass or fail.

The increasing use of performance assessment for making licensure and certification decisions in North America acknowledges the importance of assessing more than the knowledge and reasoning skills of professionals (Kane, 1997; LaDuca, 1994). Short-answer write-in questions are probably the most common form of performance

assessment being implemented at the licensure level but they are the least authentic. Several professions use more direct samples of performance.

Certification of teaching competence (Committee on Assessment and Teacher Quality, 2001a) and airplane pilot testing (Baker & Dismukes, 2002) both use actual samples of performance or sophisticated simulations as a component of a high stakes assessment. Similarly, the Objective Structured Performance Related Examination (OSPRE®) uses simulation for assessing police officers for promotion (Centrex, 2006b).

The teaching example comes from the United States where the National Board of Professional Teaching Standards has pioneered a national, voluntary system for certifying accomplished teaching that is widely recognized (Committee on Assessment and Teacher Quality, 2001b). There are two primary components. One is a portfolio with four entries of work samples, such as an analysis of students' growth in reading and writing ability, and a video sample of the teacher's classroom work with a commentary from the teacher that reflects on the value of the exercise shown in the video. The second component is a series of written exercises based on content relevant to the teacher's subject matter. This component is completed at an assessment centre. Two assessors score each component according to a holistic rubric with four performance levels defined in language drawn from the relevant content standard (Committee on Assessment and Teacher Quality, 2001b; Wolfe & Gitomer, 2001). The raters are professional teachers who must successfully complete a training program before assuming their new assessment role.

Airline pilots may be the most scrutinized of all professionals and simulation has long been a part of their assessments. Pilots and flight crews are assessed at least annually to certify that they are competent to fly. The Line Operation Evaluation (LOE) is a

common performance-based component of these assessments (Baker & Dismukes, 2002). In a LOE, a flight crew works through a flight scenario within a simulated cockpit while a pilot instructor judges their performance on a series of events built into the scenario, such as weather problems or equipment malfunctions. The pilot instructor completes complex, descriptive scoring instruments designed to capture how well the flight crew performs. Observable behaviours such as "Crew discusses need for takeoff alternative" and "Crew discusses clearance brief items" are scored on a series of four or five-point rating scale items (Holt, Hansberger, & Boehm-Davis, 2002). In addition, crew members are scored on how well they performed relevant technical skills such as "Cockpit preparation: checklist, W&B, performance." Finally, there are global ratings for both individual and overall crew performance. Pilot instructors receive considerable training to ensure that they understand the criteria for scoring and that they can score consistently.

The OPSRE® is an example of a performance-based assessment that is being used to make high stakes promotion decisions for police officers. The task component is comprised of simulated, rank-specific work tasks (Centrex, 2006b). Test takers have a 45-minute period to prepare from materials provided by the testing body. They then complete a series of seven short (five-minute) encounters with trained role players. Their performance is scored against a descriptive behavioural rating scale with items based on competencies relevant to each task. Up to five competencies may be assessed for any one task. A sample task provided to prospective test takers starts with an e-mail from an inspector regarding problems with an experienced sergeant. The test taker then meets with the "sergeant", a trained role player. Performances are rated on a scale that runs from "Superficial" through "Thorough" on five dimensions: "Effective communication",

"Maximising potential", "Planning and organizing", "Problem solving" and "Respect for diversity". Raters are senior police officers who must successfully complete a five-day training course specific to their role as an OSPRE® assessor.

These are three examples of certification assessments. For each one, the sample of tasks is relatively small but each task is an extended, complex sample of performance. Although the OSPRE® encounters are short, they are part of a larger whole that includes the 45-minute preparation component. The scoring instruments for all three examples are variations on rating scales and rubrics representing different dimensions of a competency construct. The instruments are also generic, meaning the items or categories within each one are not directly linked to the content of each performance exercise. Instead the units within each instrument represent more global aspects of a construct, for instance the "Maximizing Potential" item from the OSPRE® example.

In all three of the above examples, raters receive extensive training. The goal of the training is to establish a common understanding of the criteria for judging performance within the rater pool, thereby limiting bias, and to minimize the random error that results when raters incorrectly complete scoring instruments. The tasks are authentic, complex samples or simulations from the professional domain. Reliable scores are achieved through the use of carefully constructed rubrics and rater training. In these examples, the authenticity of the tasks is high but the sample of performance is relatively small, which limits reliability. To compensate, rater training is extensive and sophisticated generic rubrics are employed.

The use of performance assessment in medicine and other health professions has taken a different form than the certification examples already described. There is a long

history of using performance assessment in medical education that includes today's extensive use of the Objective Structured Clinical Examination (OSCE) (Swanson, Norman, & Linn, 1995). The OSCE is used for assessing performance at all levels of medical training, including the licensure level. Licensure assessments in the health professions, including medicine (Grand'Maison, Lescop, Rainsberry, & Brailovsky, 1992; Reznick, Blackmore, Dauphinee, Smee, & Rothman, 1997), physiotherapy (Parker-Taillon et al., 1992), pharmacy (Austin et al., 2003), and most recently, nurse practitioners (A. Roots, British Columbia Nurses Association, Personal Communication, March 3, 2006), all have an OSCE component.

In an OSCE, test takers rotate through a series of stations. Each station is based on a specific patient case, as portrayed by a standardized patient. The standardized patient is a lay person trained to present a patient problem reliably and realistically. (Occasionally a manikin or other form of simulation is used instead of a standardized patient.) At each station test takers are expected to perform a specified task, such as take a history, with the standardized patient. The purpose of the OSCE is to objectively measure clinical competence across a range of patient cases and clinical skills such as history taking, physical diagnosis, doctor-patient communication, and patient management. Scoring may be done by the standardized patient after each encounter or it may be done by an observer, usually a health care professional.

When the OSCE was first introduced over 30 years ago (Harden et al., 1975), assessments of clinical competence were less objective. Medical trainees were assessed on their performance with a random selection of one or two patients. Scoring standards were largely determined by the one or two physicians who acted as examiners. With the

OSCE, patient problems are pre-selected and checklists specific to the patient problem and task are used for scoring. The use of case-specific checklists introduced standardized scoring criteria to the assessment of clinical competence and this characteristic has been essential to the OSCE gaining widespread acceptance.

With the OSCE format, a broad sample of patient cases and tasks is taken but each case-task combination is understood to be but a small sample of performance. The cases are more natural than items on a standardized test but more controlled than observations from a field setting, such as the classroom videos used by the teaching assessment. Using pre-determined cases allows checklists with case-specific items like "Elicits description of headache" to be developed. The use of checklists minimizes error variance due to assessor bias, as less interpretation is required for scoring. The potential for raters to make clerical errors is also reduced as the recording task is simple. The drawback is that developing reliable and valid checklists for each patient case is a time-consuming and costly effort.

The authenticity of OSCE simulations are limited by the short, fixed time given for each patient encounter and by the limits of human simulation. When the time allowed is short, tasks become very specific, such as "Conduct a relevant physical examination". This is a less authentic task than asking the test taker to assess the patient, which would include taking a history, and an even less authentic task than asking the test taker to manage the patient, which could include ordering investigative tests and making treatment decisions. There are also limits to the physical symptoms that can be simulated by the standardized patients. Symptoms like pain, shortness of breath, limited range of motion, and injuries such as abrasions and wounds, can be presented realistically.

However, symptoms like a hot and swollen knee cannot be simulated. The short encounters allow for more samples of performance and for more raters to contribute to the assessment. With the OSCE, a degree of authenticity is traded for the improved reliability that can be attained with a larger sample of performance, a larger number of raters and detailed scoring instruments.

The most significant differences among these four examples of performance assessment are in the methods used to control for error variance. The reliability of the certification assessments (i.e., LOE, OSPRE®, and teaching certification) is enhanced by the use of generic, largely holistic instruments and extensive assessor training while the reliability of the OSCE depends more upon task-specific, detailed instruments and a greater number of performance samples: More measures lead to more reliable scores (Arnold, 1996). OSCE-based rater training, when physicians observe and score, is limited.

The difference between the two approaches can be attributed to the differences in their purpose and to the logistical constraints they face. With certification, the intent is to assess the competency expected of experienced professionals. Complex samples of performance are required for the assessments to elicit discriminating levels of performance among experienced professionals. Practical constraints limit the number of samples of complex performances that can be obtained for each assessment. Certification of excellence may be voluntary, as with the teaching certification in the United States or based on individual career paths, as for pilots and police officers. In either case, the assessments do not need to be administered at a fixed point in the year. Continuous intake and flexible scheduling parameters permit year-round administration of assessments. The

combination of these two factors allows time for complex performance samples to be gathered and makes rater training a worthwhile investment, as raters can be expected to participate in series of assessments over time.

When assessments are used for licensure decisions, they are typically administered at the point of graduation from a training program. An extended testing period would unfairly delay entry to practice for some test takers. For example, medical licensure examinations in Canada are offered twice yearly. Rather than investing heavily in rater training (since the same individuals may not be available from one administration to the next), Canadian medical licensure organizations have invested in developing easily scored instruments and maximizing the number of tasks. This approach is arguably more acceptable for licensure than for advanced certification because the level of competency being assessed is that required for entry into practice, not the higher levels of professional achievement usually required for certification and promotion.

When Harden (1975) first introduced the OSCE, a critical component of the rationale for its use was its reliance on checklists. The checklists made explicit the performance criteria for each task within the context of a specific patient case and therefore the OSCE produced reliable scores. The OSCE was an "antidote" for the subjectivity inherent to traditional oral examinations and for the subjectivity ascribed to the rating scales used to assess medical students during their clinical rotations (Streiner, 1985).

However, as the OSCE gained acceptance as a way to reliably assess medical students (van der Vleuten & Swanson, 1990), attention started to shift to scoring approaches that would be more congruent with developments in theories of clinical

expertise. While checklists are still the dominant format for scoring instruments, they may be complemented with rating scales that assess behavioural aspects of the test taker's interaction with the patient. When this occurs, one person usually scores both instruments. OSCE is a generic term for an assessment format with many variations based on length of time per station (e.g., 5 to 25-minute stations), the nature of the scoring instruments used and who scores (e.g., standardized patient by recall versus physician observer).

What follows is a synopsis of a construct of clinical expertise. Understanding this construct is a necessary precursor to the review of the studies that shaped this research. The synopsis also provides part of the framework for the research questions.

Clinical Expertise: An Evolving Construct

Ericsson and Charness (1994, p. 731) defined "expert performance as consistently superior performance on a specified set of representative tasks for the domain that can be administered to any subject" (p. 731). From these criteria, clinical expertise can be defined as the consistently superior ability to provide effective medical care to patients that is the result of knowledge, skills and judgment acquired through formal education and experience in the practice of medicine.

Early assumptions about clinical expertise assumed that skills such as problem solving, history taking and physical examination could be acquired and then applied across patient problems. However, there is a low correlation for physician performance across patient problems, meaning performance with one patient is a poor predictor of performance with the next one. This finding effectively undermined the validity of skill-based constructs of clinical expertise. The phenomenon was labeled content specificity

(Elstein, Shulman, & Sprafka, 1978) and the implications for performance-based testing were significant. A reliable measure could only be achieved if performance was tested across a broad sample of patient cases (Kane, Crooks, & Cohen, 1999; Miller & Linn, 2000; Swanson et al., 1995; Wolfe & Gitomer, 2001). In North American medical education one of the effects was a decline in the use of performance assessments like the traditional oral examination, which were based on only one or two patient cases (Norcini, 2002), and an increase in the use of the OSCE which sampled clinical competence over a broader range of patient cases (Swanson & Norcini, 1989).

The perception of clinical expertise as a problem solving construct has given way to more complex hypotheses rooted in cognitive theory about the kinds of knowledge that lead to clinical expertise (Norman, 2005b; Schmidt, Norman, & Boshuizen, 1990). Clinical reasoning is a marker for clinical expertise and is defined by Schmidt and his colleagues (1990) as the cognitive aspect of the ability to assess a patient problem, arrive at a diagnosis and generate a treatment plan. Clinical reasoning is what enables a physician to judge wisely the actions required in a specific context and is a function of a physician's ability to integrate the formal knowledge acquired through considerable study with experiential knowledge gained through extensive caring for patients.

Norman (2005b) categorizes the formal knowledge acquired by physicians into two broad categories: Basic biomedical science and disease-related schemas (i.e., matrices of disease signs and symptoms weighed against diagnostic hypotheses) and adds a third one, experiential knowledge. This third category refers to the repertoire of mental representations or exemplars of the different patient problems that a physician constructs from experience with patients. When faced with a patient problem, the expert physician

has access to a complex knowledge base that encompasses all these categories. The expert physician may utilize rapid, non-analytic pattern recognition based on an exemplar patient presentation drawn from experience to generate a diagnostic hypothesis or she may rely on a sophisticated analysis of diagnostic hypotheses (Elstein, 1993) based on a disease schema or she may draw on basic science concepts. What strategies the expert physician employs depends on the nature of patient problem relative to her knowledge and experience. For a problem the physician sees frequently, a diagnostic hypothesis may be reached without any analytic thought, solely because it is something that has been seen many times. For infrequent problems or ones with rare complications, the physician may draw on schemas that seem related or she may invoke base science concepts to resolve the diagnostic challenge (Schmidt et al., 1990). With expertise comes efficiency and efficacy. Even when solving unfamiliar problems, clinical experts screen out irrelevant information more quickly as they generate and rule out diagnostic or treatment hypotheses. As a consequence, each physician has an idiosyncratic approach to each patient problem that evolves over time. From an assessment perspective, this is a complication.

Content specificity was hypothesized to be measurement error caused by unidentifiable dissimilarities of content. Measurement error was responsible for the low correlation of performance across cases. However, now there is evidence suggesting that content specificity is more likely a result of physicians selecting different problem-solving strategies (Eva, Neville, & Norman, 1998) than it is an effect of content differences. If the differences in a test taker's performance across patient problems are a result of the test taker employing different strategies for different patient problems, then

test taker-problem interaction may be construct-relevant variance. Viewed this way, the difference in performance is an indicator of the test taker's expertise relative to each patient problem and therefore is not measurement error. In the same vein, Nichols and Sugrue (1999) argued that implicit simplistic cognitive assumptions lead to instrument designs that fail to represent the test taker-problem variance predicted by complex constructs like clinical reasoning, thereby undermining reliability analyses. Determining when test taker-problem variance is relative error and when it is construct-relevant variance is the challenge.

As thinking about medical expertise has moved from a skill-based problem-solving construct to a clinical reasoning construct, skill-based checklists have been increasingly criticized. The list of negative consequences includes promoting rote performances based on memorization of checklist criteria (as students inevitably discover their contents) (Miller & Legg, 1993), trivializing the role of expert observers to being simple recorders of observations (van Luijk & van der Vleuten, 1992), and rewarding test takers who use the "shotgun" approach of peppering the patient with questions or physical examination manoeuvres with little regard for their relevance to the patient problem (Reznick et al., 1998). These criticisms are similar to the criticisms levelled at large-scale standardized tests, particularly those related to narrowing of the curriculum and rote learning. An implicit assumption underlying the introduction of checklists to clinical assessments was that task-based thoroughness in gathering data from the patient (by taking a history or completing a physical examination) was a hallmark of clinical expertise. However, there is no theoretical basis for this assumption as a detailed or thorough approach to the patient has not been linked to clinical expertise or to diagnostic

accuracy (Cunnington, Neville, & Norman, 1997; Norman, van der Vleuten, & De Graaff, 1991; van der Vleuten, Norman, & De Graffe, 1991).

In Norman's (2005b) view, the implication of current expertise theory for the assessment of clinical competency is clear. Since there are many ways to approach each problem and each physician's approach will be a reflection of her own schemas and exemplars, the most defensible approach to assessment is to measure the number of successful outcomes each test taker achieves across as many short patient problems as possible. He argues strongly against the use of checklists for performance assessment, as there is no theoretical or empirical support for their use (Norman, 2005a). Rating scales which do not detail how the test taker reaches a diagnosis or treatment plan fit better with clinical reasoning theory than do checklists and as discussed in the next part, have empirical support.

Scoring Instruments: An Empirical Perspective

As medical expertise theory has moved away from assumptions about thoroughness, rating scales have regained the attention of medical educators, especially when used within the structured environment of an OSCE. Since the 1980s there have been many studies that compared generic rating scales to task-specific checklists within the context of the OSCE. However, most of these studies could not directly compare the two instrument formats as the rater who scored the checklist also scored the generic rating scale (e.g., Hodges et al., 1999; Regehr, MacRae, Reznick, & Szalay, 1998; van Luijk & van der Vleuten, 1992). When the same rater scores both instruments, any interpretation of the generic rating scale score is confounded by the possible influence of the checklist criteria on the rater's judgment and the potential for the checklist criteria to

standardize the rating scale scores across raters. In an early study (van Luijk & van der Vleuten, 1992) that showed no significant difference in the overall reliability between generic rating scale scores and checklist scores, this issue was acknowledged. Although seen as a limitation to the study, there was some indication from a sub-sample of their data that the checklists did not have a significant effect on the rating data. While reassuring, it was not conclusive evidence.

The finding that generic rating scale scores were more reliable was reported again in a study by Cunnington and his colleagues (1997), who made a more direct comparison between the two instrument formats. The raters who scored the generic rating scale in this study did not score the task-specific checklist. However, these raters were given the list of items from the checklist as a reference; so the major difference in relation to the previous study was the raters were not required to actually score the checklist. Although no explanation was given, providing the raters with the checklist items suggests the researchers were reluctant to allow them to score a generic rating scale without some form of case-specific criteria.

Four more recent studies have examined the same issue and reported similar findings. The first of these four (Reznick et al., 1998) used data collected from 372 participants during the 1997 administration of the licensure OSCE administered by the Medical Council of Canada. Scores from a seven-item generic rating scale were compared to case-specific checklists across four patient cases. The raters were physicians and they scored either the rating scale or the checklist, not both. The generic rating scale focused on clinical behaviours deemed to be dimensions of clinical expertise and was

comprised of items for history taking skills, physical examination skills, management skills, knowledge, logic and flow, communication, and professional behaviour.

For three of the four cases the generic rating scale scores performed as well as or better than checklist scores based on their item-total correlations (ITC) but for one case the ITC for the generic rating scale scores was lower. Low correlations between the rating scale scores and the checklist scores and generally low agreement on pass-fail decisions were also reported. The one common measure between the two sets of raters was a six-point holistic rating item on the test taker's overall approach to the patient with anchors that ran from Inferior to Excellent. The inter-rater correlations for this item had a large range (.17 to .64) and it was not used to control for rater effect.

The conclusion was that the two instrument formats were measuring different constructs or they were measuring the same construct with considerable error. Either way, the two formats were not interchangeable. More research would be needed to determine, for instance, whether differences were an instrument effect or a construct issue.

In a second study (Regehr et al., 1998) checklist scores on their own, checklist scores followed by a generic rating scale score, and an independent generic rating scale score were compared. In this study the data were from a surgical OSCE. All 53 test takers were surgical trainees and there were no standardized patients. Each station in the OSCE assessed a surgical technique as performed on a model. All the raters were surgeons. For each station, one rater scored the checklist and a generic rating scale and the other scored only a generic rating scale. The rating scale had seven dimensions related to operative performance; for example, "respect for tissue", "time and motion", "knowledge of

instruments" and "knowledge of procedure". The checklists were specific to each station and had 20 to 40 items related to the specific actions required for each procedure.

There was no significant difference in the internal consistency of the total OSCE score, whether based on the checklists or on the generic rating scale. As is expected when scores from two instruments are measures of the same domain of performance, the combination of the checklist scores with the generic rating scale scores was significantly more reliable than the checklist scores alone. However, the generic rating scale scores were more accurate at predicting residents' training level (i.e., stronger evidence of construct validity) and showed a stronger correlation with scores from an independent assessment of the quality of the finished tasks (i.e., stronger evidence of concurrent validity).

The conclusion was that checklists should not be considered an improved approach to scoring and that primary consideration should go to using expert raters and carefully constructed generic ratings instead. However, the generalizability of this study is limited by its surgical focus, as acknowledged by the researchers. This was a discipline-specific assessment focused on technical skills, free of the variance in performance associated with patient interactions. Although the raters were not trained as such, a certain degree of standardization was achieved by only using surgeons to score and the technical nature of the tasks further constrained the scope of judgment required of them, minimizing the potential for rater error.

In a third study, Hodges and his colleagues (1999) compared the relationship of checklist scores and generic rating scale scores to test taker expertise and diagnostic accuracy. The study used two 15-minute history-taking cases based on patients with

common psychiatric complaints. Physicians were the raters. The 42 participants were clinical clerks (novices), residents (intermediate) and practicing physicians (expert). The raters scored a 22-item checklist and a generic rating scale with five five-point items for each of "knowledge and skills", "empathy", "coherence", "verbal expression" and "non-verbal expression".

The results showed practicing physicians had the highest scores with the generic rating scale and the lowest scores with the checklists, which led the researchers to conclude that using checklists penalizes more experienced physicians by rewarding thoroughness as the checklist scores did not distinguish between the different levels of expertise. However, the researchers were unclear whether the rating scale scores would be sensitive to increasing levels of expertise. The rating scale scores did not distinguish any more accurately than the checklist scores between clinical clerks and residents who were closer together vis-à-vis their level of training than were residents and practicing physicians.

An unacknowledged limitation to this study was the nature of the checklists. They were carefully constructed for a clerkship (novice) OSCE and it might be argued that in fact the checklist scores did distinguish between levels of expertise, just not in the way that was anticipated by the researchers. The target group (clerks) had the highest mean checklist score and the mean checklist scores *decreased* as level of expertise increased, with two different cohorts of participants. The expected pattern of scores *increasing* with expertise might have occurred if the checklists had been constructed to assess physicians in practice. Using checklists developed to assess novices to assess experts leaves open the question of whether the effect was one of instrument format or content. As with some of

the earlier studies, the same rater scored both the checklist and the generic rating scale, further confounding the interpretation of the results.

In a fourth study by Hodges and McIlroy (2003) checklists were again compared to a generic rating scale based on their sensitivity to levels of training. Each rater scored two instruments; a task-specific checklist and a generic rating scale comprised of five-point items for each of "empathy", "coherence", "verbal communication" and "non-verbal expression". Unlike earlier studies that were explicitly or implicitly considering rating scales as replacements for checklists, this study focused on the validity of augmenting checklist scores with rating scale scores. The data were from ten cases that sampled patient problems from the major medical disciplines and the raters were physicians. The 57 participants were third and fourth year medical students. The total rating scale score was more internally consistent than the total checklist score (i.e., alpha (α) coefficient of .70 versus .54) and there was evidence of significant construct validity for the rating scale scores based on mean score differences for the two groups of participants. The difference in rating scale mean scores for the two groups was largely attributable to two of the four rating items. Senior students scored significantly higher on coherence and questioning. Hodges and McIlroy made the point that experience in clinical settings should impact positively on medical students and that their ability to carry out a coherent interview marked by logical questioning should improve with practice; more so than say empathy. The conclusion was that identifying which items within a rating scale lead to more accurate and consistent scores is as important as assessing the generic rating scale as a whole. Whether or not scoring the checklists had an

influence on the ratings was not addressed and was not as relevant in this study as the rating scale was being assessed as a way to augment checklists.

The specific content of the checklists (Hodges & McIlroy, 2003a) was not described but may be inferred from the descriptor "task-specific" to be a series of items identifying the critical steps to be taken or information to be gathered during each patient encounter. The rating scale assessed the quality of the interaction with the patient without any items directly related to the quality of the case-specific information gathered. In this study, the dimensions being assessed by the two formats were explicitly different.

The assumption of most of these studies was if rating scales produced more reliable scores, then they could replace checklists, so long as they were at least equally valid. However, in every study the rating scales were generic and skill-based with one item for each dimension of a broad, largely skill-based construct of clinical expertise. Generic instruments like these offer the greatest gain in utility as creating one instrument for scoring all patient cases would greatly reduce the time and cost of developing case-specific instruments and might simplify rater training issues. However, these studies do not address whether the generic rating scales are measuring the dimensions of clinical expertise they were intended to measure nor do they resolve whether the reliability of the scores is at all dependent on raters' exposure to the case-specific checklists.

These same issues examined within the context of a different competency construct led to a different conclusion. A study led by Reilly (1990) with data gathered from group leadership exercises administered at an assessment center also compared checklist scores to rating scale scores, although the comparison was not strictly the intent of the study. The data were from two group exercises that were scored with descriptive

rating scales by raters trained to score on eight dimensions: "Teamwork/interpersonal skills", "leadership", "problem solving", "work orientation", "attention to detail", "comprehending/following instructions", "planning and organizing" and "safety awareness". The group tasks were assembly problems. For example, in one exercise the participants were asked to plan and organize an approach for building a robot. Each rater scored two of the participants. With the first group of participants, the raters used the rating scales, as always. Then they participated in an exercise to develop checklists comprised of items based on the observed behaviours that they had identified as key to making their own rating decisions. Each checklist item would be scored as not done, done once, or done more than once. The items were organized according to the eight dimensions of the rating scale. After the raters observed a second group of participants, they immediately completed the checklists to aid their recall; then they scored the rating scale.

The results showed a curvilinear relationship between the number of items on the checklist for a dimension and the correlation between the checklist scores and the rating scale scores for that dimension, meaning that score correlations increased as the number of items increased until approximately 12 items were reached. After 12 items, adding further items did not improve the relationship. The correlations were especially strong for dimensions where there had been adequate opportunity for the participants to demonstrate the behaviour being assessed. The most important result for the researchers was that the use of the checklists doubled the convergent validity of the ratings (based on the correlation of dimension scores across two simulation exercises with the same participants). Although the assessors' ability to use the rating scales may have been

improved by the experience gained from their participation in checklist development, it was noted that prior studies in assessment centres had not reported an experience effect on convergent validity indicators. The researchers believed that a more likely explanation was that the checklists reduced the cognitive demand on assessors by focusing them on specific sets of relevant behaviour, by enriching their recall, and by categorizing what they recalled along the same dimensions as the rating scale.

Almost as an aside, it was noted that the indicators for convergent and discriminant validities were slightly higher for the checklists than for the rating scales. This result led the researchers to suggest that because the checklists are easier to score and more objective, and given that using them instead of rating scales would simplify rater training and provide more specific feedback to the participants, that checklists might be preferable. They speculated that the raters were being subtly influenced by participant behaviours that were related to doing well in the corporate culture but were not necessarily related to effective task performance. If so, the slightly higher validity for the checklists may have been a result of minimizing this effect.

The comments from Reilly's group (1990) echo the arguments made for OSCE checklists 15 years earlier and run counter to current arguments for the use of rating scales. An interesting difference between the checklists used by Reilly and his colleagues and OSCE checklists is how they were developed. In Reilly's study the checklist items were derived from the raters' reflections on what behaviours had had an important impact on their rating decisions. OSCE checklists are commonly developed by content experts and the items reflect their judgment of what test takers should do. Whether or not these

are the same criteria the content experts would invoke if they were asked to holistically score a series of test takers first is unknown.

Reilly and his colleagues (1990) went further and considered the impact that the cognitive challenge of scoring complex performances with a limited number of items might have on the raters. Their hypothesis was that too high a cognitive challenge would lead raters to find internal rules that simplified scoring judgments, a process that could lead them to incorporate unintended criteria that were not directly related to the construct being measured.

A related concern was raised by Hunter (2001) based on a large scale study that compared holistic to analytic scoring approaches for scoring a writing skills assessment. Two of the recommendations from the Hunter study are of particular relevance here. First, while the holistic scores were quite robust for program evaluation, caution in employing holistic results to summative assessments was urged until the differences between *what* the two approaches were measuring was better understood. Second, Hunter argued for a theory-based approach to rater training that would ameliorate the personal and idiosyncratic constructs that raters bring to a holistic scoring task.

Concern about the dynamic between rating scales and raters was addressed from yet another perspective by Marbry (1999) based on her analysis of the scoring rubrics commonly used for scoring assessments of writing ability. She argues that these rubrics are generally not as holistic as proclaimed. The strength of the rubrics lies in the rich explanations provided to guide the rater in making decisions on each dimension. According to Marbry, their weakness is the limited number of dimensions. Scoring with only a few dimensions creates a narrow perspective on writing ability, which is a

complex or "wide" construct. In her view, rubrics improve inter-rater reliability by limiting the raters' scoring choices to a few dimensions with a limited number of performance levels for each dimension.

Restricting rater choices lessens the risk of disagreement. Rater training further narrows the focus of the raters, leading Marbry (1999) to make the observation that communal tunnel vision should not be mistaken for informed consensus about the quality of performance. Part of her concern stems from the tendency of many rubrics to emphasize low level criteria like spelling and organization over the more important qualities of persuasiveness or figurative use of language. The lower criteria are easier to measure reliably as less rater judgment is required, leading to less error. While such rubrics may promote score reliability by standardizing scoring criteria, she argues that they also standardize the teaching of writing with negative consequences for the learning of creative writing ability. According to Marbry, using generic rubrics redefines the construct from writing ability to the ability to write to the rubric and she cautions that an emphasis on reliable scores leads to a loss of content relevance, which is a loss of validity.

Rating scales, rubrics, checklists, and holistic judgments have all been used to score performance assessments. Nichols and Sugrue (1999) argue these instruments are all vulnerable to under-representing the complexity of observed performance, minimizing the differences between test takers and undermining the validity of decisions. Their argument is based on examples from a large-scale assessment of constructs of mathematical ability. They describe how complex constructs may be simplified at any step in the test development process, but especially when simple trait-based theories

implicitly drive the design of scoring instruments. They contend that the move to educational constructs defined by multiple cognitive processes is not being matched by parallel developments in test design. The tendency to use unidimensional instruments restricts how well the variance inherent to these multidimensional constructs is represented and generates scores that may not discriminate between individuals as effectively as they should.

Scoring instruments can be placed along the continuum of scoring strategies developed by Hunter and his colleagues (1996). At one extreme is "general impression" scoring (the most holistic) and at the other extreme there is "atomistic" scoring (the most analytic). This continuum mirrors the differences between cognitive and behavioural assumptions about learning. Cognitive theory sees learning as a constructive process mediated by context and prior learning whereas behavioural theory views learning as an atomistic process of adding up bits of knowledge. More holistic instruments, like rubrics and rating scales, represent a cognitive approach to scoring while checklists are a behavioural approach to the same task.

As you move from general impression scoring, with no explicit criteria for judging the overall product or performance, along the continuum toward more analytic approaches, scoring becomes more an inventory of performance and there are increasing constraints on the degree of judgment that assessors can exercise. More holistic scoring methods rely on rubrics which define performance criteria broadly and provide descriptors for each level of achievement within a criterion. These instruments allow judges to interpret performance within the given guidelines and thereby allow the judges to take into consideration more of the performance, at least in theory. Marbry (1999)

would argue otherwise. As scoring becomes more analytic, scores become the result of summing scores from a set of rating scale items and simpler rubrics or rating scales are more commonly employed. There is less interpretation required from the judges and the construct is more narrowly defined. At the analytic extreme, checklists that decompose the task into a series of dichotomously scored steps become the most common tools for scoring. There is little interpretation required of the judges and little consideration of process. In essence, behavioural scoring is imposed on a cognitive task.

The argument against checklists for scoring professional competencies is relatively simple. Checklists are too detailed for assessing developing expertise. The reliability and objectivity associated with checklist scores is offset by their limited discrimination across levels of expertise. Given this argument, the case for using some form of rating scale is easy to make. For example, analytic rating scales have some of the objectivity of checklists and when used in the structured context of an OSCE, produce very reliable scores that discriminate across levels of expertise. Using generic rating scales adds a level of utility that case-specific checklists can never provide.

The strongest argument for generic rating scales is the reliability claim. The validity evidence is not overwhelming. For instance, one generic rating scale only discriminated well between residents and practicing doctors, not between clerks and residents (Hodges et al., 1999), where it did no better than the checklists. There was likely a more significant difference in the expertise of the practicing doctors as compared to the residents (or at least in the patina of expertise that comes with more years of seeing patients) than there was in the expertise of the residents as compared to the clerks, although not a difference captured by the checklists. Rating scales did discriminate well

between test takers based on correlations with the finished products of a surgical skills assessment (Regehr et al., 1998) but this result is not very generalizable to the broader construct of clinical expertise because skill-based rating scales were used to assess psychomotor surgical skills only. This is a narrow dimension of the clinical expertise construct with standardization built into the tasks, which limits inferences to the wider domain of patient care.

An outstanding concern is the lack of clarity regarding exactly what dimension of a construct is being assessed by generic rating scales, whether the evidence comes from studies on the assessment of clinical expertise, writing ability or leadership skills. Despite the use of descriptors for the rating scale items, it is not clear what criteria raters bring to bear when they make rating decisions. If score reliability is achieved by narrowing the raters' focus (Marbry, 1999), then there is a potential loss of content relevance. This is a distinct possibility with the generic rating scales that have been used to score OSCE as these instruments are not case-based and do not have a strong theoretical base. The items in the generic rating scales in the above studies were largely skill-based and emphasized the quality of the relationship to the patient, with no link to the key features dimension of each patient's problem. Key features are those critical steps or elements of a patient case that must be recognized and followed if good patient care is to ensue (Farmer & Page, 2005; Page & Bordage, 1995). Judgment, the clinical reasoning based on knowledge and experience, is required to identify the key features of a patient problem. A key features approach is used to assess clinical decision-making in written and computer-based tests (Farmer & Page, 2005) and the principles of developing these types of tests should be applicable to performance assessments. Although OSCE checklists have not been

designed along key features principles, the items within these checklists commonly include key features of the patient problem. How well any one checklist score represents the key features dimension is the question.

While generic rating scales produced reliable scores, the validity criteria were not clearly defined. On what basis did the raters discriminate between levels of training? The strongest evidence of construct validity came from the surgical skills assessment (Regehr et al., 1998) and it was the least generalizable. Other limits to these studies, as recognized by the researchers, included lack of a common instrument for assessing interrater reliability (Reznick et al., 1998), one rater scoring both instruments and data based on two psychiatric cases (Hodges et al., 1999), and one rater scoring both instruments with a generic rating scale that was really focused on the quality of the interaction with the patient, not on the broader construct of clinical expertise (Hodges & McIlroy, 2003).

From studies focused on methodology (i.e., rating scales versus checklists) there is evidence that checklist scores do not discriminate sufficiently between levels of training. Generic rating scale scores discriminate better between levels of training but the validity evidence does not clarify on what basis. Hodges and McIlroy (2003) found that the two best items within their generic rating scale were coherence and questioning (verbal behaviour) and emphasized the importance of validating items within the instrument; a reminder that format is not everything.

If the strengths of the two methodologies were combined, would a more valid instrument result? Analytic rating scales better exploit expert raters than do checklists and produce reliable scores. Checklists are case-specific. If the items within a rating scale are anchored by descriptors from the key features of a patient case, sacrificing the utility of a

generic instrument, then there is a stronger theoretical base for the scores because the potential to capture the key features dimension should be greater. Alternately, if the items within a rating scale are anchored with descriptors of a specific skill then the rating scale retains some of the utility of a generic scale and has the advantage of using several items to generate a score for one dimension, such as history-taking. This is different from a generic rating scale that has one item for each of several dimensions, like history-taking, physical examination skill, knowledge, and organization. A skill-based rating scale with multiple measures of one skill would represent a compromise between expertise theory and measurement criteria. With either approach, sacrificing the utility of a truly generic rating scale suggests that more construct valid approaches to rating scales can be devised, which led to the two research questions in the next section.

Conceptual Framework

This research follows from the assumption that performance assessments should be an important component to making high stakes decisions about professional competencies, an assumption that is in keeping with current learning theory and trends in educational assessment. Performance assessments allow us to have some measure of the complex skills and judgments that we expect the professionals whom we trust to teach, care for and protect us to possess. This research starts from the outstanding challenge of how to design the best possible performance assessment for any given profession, which in the health professions currently translates into how to design the best possible OSCE, best meaning it leads to the most valid and reliable scores.

The OSCE is paradoxical. When introduced over thirty years ago the OSCE was an innovative approach to performance-based assessment of clinical expertise. The

development of the OSCE was concurrent with cognitive learning theory's growing ascendance over behavioural theory and new insights into the nature of clinical expertise, in particular the problem of content specificity. The OSCE overcame this problem by sampling performance across multiple patient problems. Despite its emphasis on the use of authentic tasks, the OSCE reliance on checklists emphasized a thorough, step-by-step approach to a patient problem and deliberately limited the judgments of the raters to achieve reliable, objective scores. One way to improve an OSCE is to resolve the inconsistency of scoring complex performances with detailed, behavioural checklists.

Rating scales are an obvious alternative to checklists because scales focus the raters more on the qualitative aspects of observed performance than on the quantification of task performance. What prompted this research was the discrepancy between the empirical evidence for the use of *generic* rating scales that focused on test taker behaviour and the importance that clinical reasoning theory places on case-based criteria for assessment. Norman (2005a) argues that outcomes like diagnostic accuracy and sound treatment decisions are probably the most reliable markers of clinical reasoning because there is no one process that can be identified as best and therefore assessed. This conclusion is compatible with a key features approach which also focuses on clinical decision-making rather than on any one process. However, generic rating scales comprised largely of items such as "empathy" and "coherence" (Hodges et al., 1999) do not incorporate case-specific key features as scoring criteria. The lack of a key feature criterion may explain why the strongest evidence in support of generic rating scales is their reliability and utility. Evidence of their content and construct validity is more limited. Moving from case-specific checklists that focused on actions taken or questions

asked by the test taker to a generic rating scale that focused on test taker behaviour relative to the patient was a leap to an idealistic solution. This research takes a pragmatic step back from the ideal to incorporate more concrete scoring criteria.

The "step back" is illustrated in Figure 2. The leap from the case-specific checklists (on the left) to the generic rating scale (on the right) was motivated by clinical reasoning theory but skipped over other possibilities. Two of those alternatives are considered in this research, one based on case-specific criteria and the other on skill specific criteria (as shown with the middle boxes). Both approaches were influenced by clinical reasoning theory in that the anchors emphasize the test taker's clinical focus or judgment, not the thoroughness of the approach to the patient problem. Using a skill-based rating scale retains elements of a behavioural model (hence the dashed line from the checklist to the skill-specific rating scale) but still retains the utility of being generic to a range of patient cases. The case-specific rating scale represents the greatest loss of utility (relative to a truly generic rating scale) but has the most theoretical support from clinical reasoning theory (Norman, 2005b) because it focuses on the test taker's ability to single out clinically relevant aspects of the patient's problem.

The lower portion of the figure presents factors to be considered when comparing scores and pass-fail decisions (shown in the boxes with broken borders) as well as the criteria for making the comparisons (boxes on the right). For example, a cohort effect, perhaps due to a disproportionate number of high performers, would limit score variance and make instrument effects harder to discern.

Meaningful comparison of pass-fail decisions means accounting for case effects and rater effects. A lack of difference between the pass-fail decisions is only meaningful

if case variables are not masking an instrument effect. For instance, there would be little difference in outcomes if a case proved either too hard or too easy, which would indicate a poorly discriminating case rather than an absence of instrument effect. If there is no difference in the pass-fail decisions and there is no case effect, then psychometric criteria and utility determine whether checklist or rating scale scores are more valid, based primarily on greater discrimination.

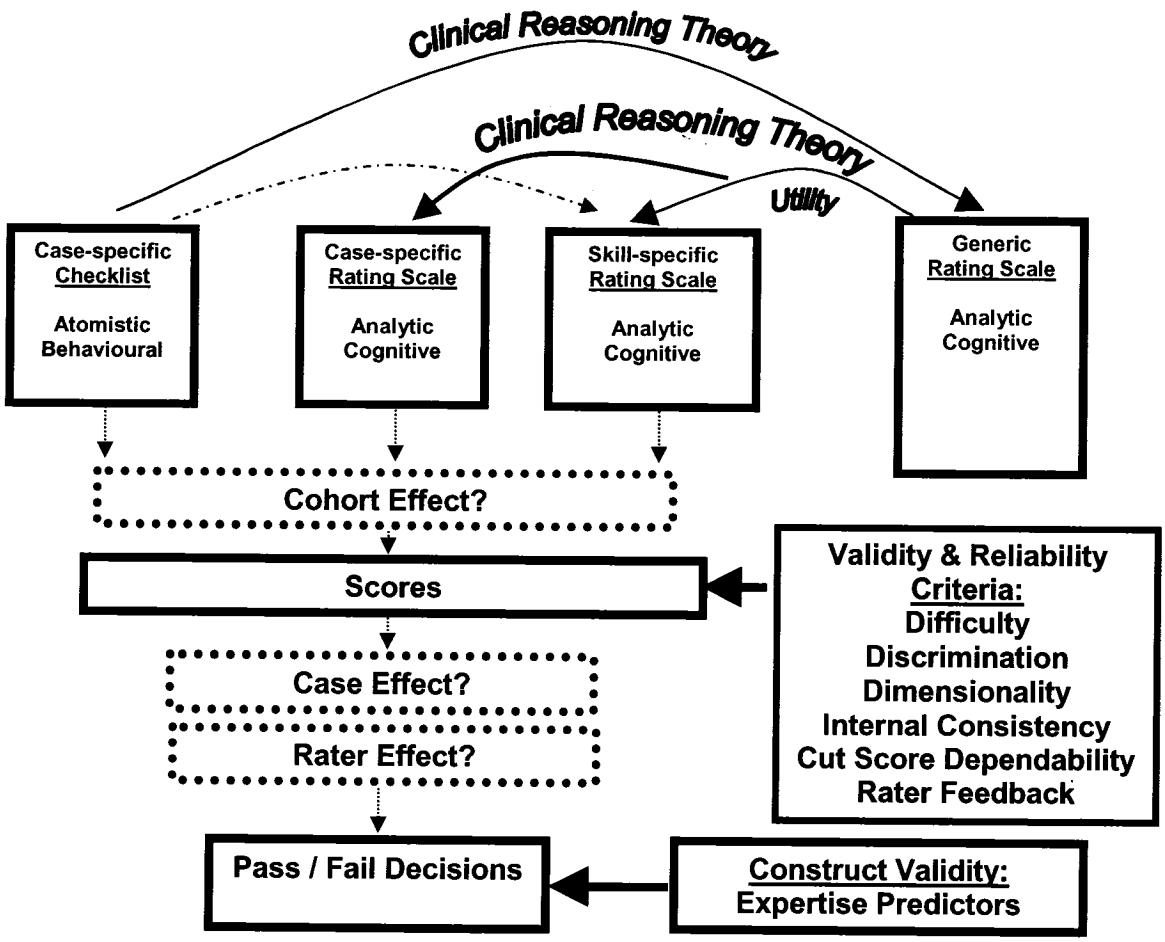


Figure 2 Conceptual framework for comparing scoring instruments

. A difference in pass-fail decisions between the rating scales and checklists could indicate that one of the scoring instruments is more construct valid, if the difference can be attributed to one instrument discriminating better between differences in test takers' levels of expertise rather than to a rater effect. If rater effect (e.g., bias, scoring errors) is not an issue then measurement criteria and the evidence of construct validity determine whether checklists or rating scales are better. Utility could still be an issue but since all the instruments are feasible, evidence of construct validity would outweigh utility.

My perception is that a case-specific rating scale represents the best compromise between utility and construct validity, while still representing progress toward a more valid or discriminating scoring instrument. My viewpoint owes much to the many discussions I have had with test takers who have failed a medical licensure OSCE, the same performance assessment used for this research. What I have heard from them suggests that generic rating scales will not represent their performance well and for some test takers may in fact inflate their scores. When asked how they approached each patient problem, their answers often suggest that their poor performance was linked to how accurately they grasped the key features of each patient problem. Two common problems described by test takers are taking a rote approach to each patient or conversely, taking a narrow approach focused solely on diagnosis. In the first instance, they may carefully gather considerable information from the patient but have a limited sense of its relative importance. In the second instance, they may have the right diagnosis but have not determined if, for example, the patient has drug allergies that will interfere with treatment. In either instance, if the test taker approaches the patient with a modicum of professional confidence and efficiency while demonstrating respect and concern for the

patient, they may score very well on a generic rating scale with items like "empathy", "coherence", "non-verbal behaviour" and "verbal behaviour" (Hodges et al., 1999), even though their clinical reasoning is actually weak. Weakness in reasoning and judgment would be captured by only one item like "knowledge and skill". The ability to identify and act on critical aspects of each patient problem is a mark of expertise yet it is not clear if this criterion is sufficiently accounted for when raters use generic rating scales.

As a clarification, the construct assessed by an OSCE, labelled here as clinical expertise, is broad and includes the clinical reasoning construct. Clinical expertise also includes dimensions like the ability to interact professionally and respectfully with a patient, the ability to elicit a relevant history, the psychomotor skills required for physical examination, and the ability to apply ethical principles of practice in a given clinical situation. The dimension of most interest in this research is the ability to elicit a relevant history. Since relevance is defined within the context of a patient problem, clinical judgment is required and so the history-taking dimension can be characterized as an expression of the clinical reasoning construct. What patient information is most relevant to each physician may vary in the details but there are key features to any patient problem that should be elicited by any physician. For this reason the dimension being assessed by the checklists and the rating scales will be referred to as the key feature dimension. This is a bit inappropriate as it is using the key feature term more loosely than intended by Page and Bordage (1995) but it is appropriately descriptive and having a succinct label for this dimension aids the discussions that follow.

This research encompasses two studies, each assessing the validity (largely based on discrimination) and reliability of scores from rating scales that were a compromise

between case-specific checklists and a generic rating scale. Table 1 is a descriptive comparison of checklists, the rating scales from the two studies and a composite of rating scales used in previous studies.

Table 1 *Characteristics of four scoring formats based on one OSCE case*

Construct	Clinical Expertise: Knowledge, skills and judgment required to provide effective medical care to patients			
Patient Case	Patient presents with trouble coping at work secondary to Depression			
Task	History-taking			
	Common Practice	**Study One**	**Study Two**	**Other Research**
Instrument	**Case-specific Checklist**	**Case-specific Rating Scale**	**Skill-specific Rating Scale**	**Composite Rating Scale**
Specificity of Instrument	Unique to patient case + skill		Unique to skill	Generic
Description of Items	Procedural steps of the task	Components defined by information relative to the patient's problem	Components defined by information relative to patient problems	Components defined by a skill-based construct of clinical expertise
Specificity of Items	Concrete – atomistic: Onset, duration, aggravating factors, related symptoms, previous episodes of similar problem	Abstract – specific to patient: Appropriateness of information about problems at work, depression-related symptoms, risk of suicide	Abstract – generic to skill: Appropriateness of information about chief complaint, related symptoms, past health	Abstract-generic to the construct: History-taking, physical examination, communication with patient, knowledge
Scoring Format	Dichotomous items	Anchored rating scale items	Anchored rating scale items	Anchored rating scale items
Score Range	Done-Not Done (0-1)	Inferior-Excellent (0-5)	Inferior–Excellent (0-5)	Inferior-Excellent (0-5)
Degree of Rater Judgment	Least	Partial-Restricted	Partial-Extended	Maximum
# of items	15 to 70	6 to 9	6 to 9	6 to 9

The headings across the top of the table show how the perspective narrows as one works from the construct (clinical expertise) toward one sample (Depression) and from it to a specific task (history-taking) within that same construct. Running down the left side

of the table are the characteristics used to compare the instruments. From the comparisons come the research questions that ask which instrument is the most valid and reliable. As shown in Figure 2, validity indicators for the scores will include measures of difficulty and dimensionality. Item-total score correlations (ITC) are an indicator of discrimination and are critical to the validity comparisons. Rater feedback will also be considered. Internal consistency will be the critical reliability indicator for scores. Dependability will be the critical reliability indicator for the pass-fail decisions and the degree to which the expertise predictors in the regression model predict the decisions will be the basis for making comparisons across instruments.

Research Questions

1. Are *case-specific* rating scales better than case-specific checklists at representing professional competency? Specifically:
    a. How does the discrimination and reliability of case-specific rating scale *scores* compare to that of checklist *scores* in a high stakes performance assessment of professional competency?
    b. How does the discrimination and reliability of case-specific rating scale *pass-fail decisions* compare to that of checklist *pass-fail decisions* in a high stakes performance assessment of professional competency?
2. Are *skill-specific* rating scales better than case-specific checklists at representing professional competency? Specifically:
    a. How does the discrimination and reliability of skill-specific rating scale *scores* compare to that of checklist *scores* in a high stakes performance assessment of professional competency?
    b. How does the discrimination and reliability of skill-specific rating scale *pass-fail decisions* compare to that of checklist *pass-fail decisions* in a high stakes performance assessment of professional competency?

These questions address issues not yet explored by previous studies. Limitations of earlier studies include small participant groups, having the same examiner score both instruments and focusing on a generic instrument, even though case-specific features are important for assessing clinical reasoning. The design for these two studies offers the power of a large number of participants, independent raters linked with a common instrument and comparisons to checklists designed to assess clinical competence at a higher level of training than that of medical students. In addition to the scores, this research examines the pass-fail decisions by comparing logistic regression analyses that are based on a model of expertise and considering the dependability coefficients for the cut scores, as described further in the next section.

Study Design

Each research question is based on a comparison of rating scale outcomes (i.e., scores and pass-fail decisions) to checklist outcomes, therefore the same design is used for each study. Data are from the 2003 fall administration of a high stakes OSCE for medical licensure. Ethical approval came from the Research Ethics Board at the University of Ottawa and through the Medical Council of Canada's internal review process. The researcher was blinded to individuals' data and only group data are reported.

However, the study was run at multiple sites across Canada and the question of whether or not the ethics boards at the individual institutions should have approved the study was not clearly addressed until after the data were collected. The development and piloting of content and test scoring processes are an obligation of the Medical Council as part of their quality assurance procedures and as such do not require a research ethics review. Furthermore, the relationship between the Medical Council and the test takers is

independent of the institutions where the OSCE is administered. The alternate viewpoint is that the research was conducted within a high stakes environment for the participating trainees and as such may have fallen under the Tri-Council Policy Statement (Medical Research Council of Canada, Natural Sciences and Engineering Research Council of Canada, and Social Sciences and Humanities Research Council of Canada, 2003) on research involving humans; in particular, Article 1.1 (d). This article indicates that "Quality assurance studies, performance reviews or testing within normal educational requirements should also not be subject to REB review" and then states that "...studies related directly to assessing the performance of an organization or its employees or students, within the mandate of the organization or according to the terms and conditions of employment or training should also not be subject to REB review. However, *performance reviews or studies that contain an element of research in addition to assessment may need ethics review."* (Emphasis added.)

The conclusions of the reviewers (Drs. Nuala Kenny, Dalhousie University and Dr. Henry Dinsdale, Queen's University, Personal Communication, March 15, 2005), in part, were that that the ethical reviews that had been conducted were done so conscientiously and in good faith. They recommended that the Medical Council articulate its own explicit criteria and process for projects requiring a research ethics board review. They further recommended that the criteria should take into account the American Psychological Associate standards, the Tri-council Policy Statement and other standards in what they referred to as a new and emerging area of ethics review. In addition, they suggested that consideration be given to creating a central review board acceptable to all

the institutions collaborating in the administration of the Medical Council's examinations for dealing with similar research projects in the future.

The study design required that both studies run within one OSCE and thus avoided some of the limitations to earlier studies. The checklists for this OSCE are designed to assess medical trainees seeking entry into independent medical practice; they are not checklists designed to assess medical students. Because both the checklists and the rating scales are designed to assess the same high level of competence, any difference in the degree to which they discriminate is more attributable to differences in format rather than to differences in their purpose. The test takers were a large group (n>1,500) and all of them had a minimum of 12 months of post-graduate clinical training. As a consequence, scores needed to discriminate between test takers who were largely in an intermediate range of expertise rather than across the broader range represented by medical students, residents, and practicing physicians of at least one study (Hodges, Regehr, Hanson, & McNaughton, 1998). Although the two studies for this thesis lack the control that comes with selecting participants according to their level of expertise, they can be seen as a more authentic comparison of instruments because the data are collected during a live examination, not an administration run for research purposes only.

The study design did have to work within the logistical constraints of a specific OSCE so the two patient cases were selected from the 12 that were available. The same two cases (Depression and Delirium) were used in both studies. For the first study, case-

specific rating scales were developed for each of the two cases. The task for both cases was history-taking, so one skill-specific rating scale was developed for the second study.

In order to overcome the limitation of having the study instrument scored by the same physician as the one who scored the checklist, additional physicians were recruited to score the study instruments. Using independent raters meant that controlling for rater effect was important. Therefore the physicians recruited for the two studies also scored the same patient interaction rating scale (PIRS) as the physicians who scored with the checklists. Scores from the common instrument were used to assess the degree of rater agreement. The study physicians are referred to throughout this report as assessors and the non-study physicians are referred to as examiners. Rater is used as a generic term.

Test taker data included scores from all instruments, demographic data (i.e., age and gender) and data related to their professional training (e.g., medical specialty, scores from a knowledge-based examination and year of graduation from medical school). As well, demographic data and feedback were collected from the physicians who participated in the studies. Scores were compared statistically (e.g., mean scores, internal consistency, variance ratios and dependability of the cut scores). The discrimination of pass-fail decisions was examined by comparing the predictive accuracy of the logistic regression analyses (using a three-variable model of expertise) across instruments and by comparing the proportion of variance in the decisions explained by the analyses.

The similarity in the test questions and study design made it possible to compare the case-specific rating scales to the skill-specific rating scale and to compare rating scales to checklists. Although not explicitly part of the research questions such comparisons were seen as a way to cross-reference the results and enrich the discussion.

Methodology

To answer the two research questions, two studies were conducted, each using the same methodology. The participants were recruited and assigned to their tasks in the same way, the study instruments were developed in tandem, and the data analysis procedures were identical. Data were collected from the same two patient problems or cases, Depression and Delirium, within the same OSCE. Therefore the methodology outlined in the following sections applies to both studies. Details like number of participants are given by study and scoring instrument as needed.

Participants

Test takers

Test taker data were collated by the Medical Council and assigned random identification numbers before being released to the researcher. Only group data are reported. All test takers had a medical degree from a recognized school and had completed 12 months of post-graduate clinical training. They also had passed a knowledge-based test administered by the Medical Council. Of the initial 1,587 participants, 1,056 (66.5%) were Canadian post-graduate trainees. These test takers were enrolled in Canadian post-graduate training programs and all of them had a minimum of 16 months post-graduate training; 80 (7.6%) of them had more than six years of clinical experience. The other 531 (33.5%) were test takers from international post-graduate medical programs with clinical experience ranging from a minimum of 12 months of post-graduate training to as many as 39 years of clinical experience.

Test takers were assigned to a site based on their preference but their assignment to a study within a site was random. The distribution of test takers by training program, field of training, experience and knowledge was very similar across the two studies. Both cohorts were dominated by Canadian post-graduate trainees (68% and 65% for Studies One and Two respectively) and test takers were spread across fields of training in similar patterns.

Assessors

Assessors were the physicians who scored the study instruments. These physicians met the same standards as all the other physicians who scored this examination. In fact, most of the assessors served as examiners for a half day and then participated as an assessor for the other half day. Each assessor had at least three years experience in independent practice and was currently in active practice. Of the 190 assessors who scored the rating scales for the two studies, 172 (91%) consented to the use of their demographic data and their comments regarding the instruments. One assessor in Study One and two assessors in Study Two examined for both cases. Based on demographics and discipline, the assessors who consented to the use of their personal data were fairly evenly distributed across the two studies which broke down into four data sets (i.e., study by case). The percentage of women assessors across the four data sets ranged from 28% to 37% and the mean age for the four groups ranged from 41 to 46 years. Family medicine assessors dominated across the board, comprising 33% to 50% of each of the four groups. A large proportion of the assessors (37% to 44% for Studies One and Two respectively) had scored on four or more previous administrations of this examination. In Study One, 19% to 20% (for the Depression and Delirium cases

respectively) of the assessors had never scored previous administrations of this examination; in Study Two 26% to 31% (for the Depression and Delirium cases respectively) had never scored for it. In summary, the characteristics of the assessors were not markedly different for Study One and Study Two. No personal or demographic data were collected from the examiners and therefore they were not considered participants.

Instruments

Data from the test takers were collected from their registration forms and from six scoring instruments. Information about the assessors, including their comments on the scoring instruments, was collected with one feedback form. Examples of the scoring instruments and the feedback form are in Appendixes A to E.

The six scoring instruments were:

1. Study One – Case-specific Rating Scale-1 (CRS-1) for Depression
2. Study One – Case-specific Rating Scale-2 (CRS-2) for Delirium
3. Study Two – Skill-specific Rating Scale
   3.1. Depression (SRS-1)
   3.2. Delirium (SRS-2)
4. Both studies – Checklist-1 (CL-1) for Depression
5. Both studies – Checklist-2 (CL-2) for Delirium
6. Both studies - Patient Interaction Rating Scale (PIRS) for Depression and Delirium

One of the limitations to some earlier studies was that one physician scored both the rating scale and checklist. In this study the rating scales were scored by a different physician than the one who scored the checklist. However, a link between the physicians was needed to control for rater effect in the regression studies so that whatever differences were identified could be attributed to instrument differences and not to

examiner differences. The PIRS was used as the link. This was a generic rating scale designed to capture judgments about the interpersonal (e.g., rapport) and skill-based aspects (e.g., questioning skill) of the interaction between the test taker and the standardized patient. Therefore, assessors and examiners completed two instruments for every test taker they observed; either a study instrument or a checklist *and* the PIRS. Attached to the PIRS was a single-item, six-point holistic rating scale that asked the examiners and assessors to rate the test taker's overall approach to the patient as:

Borderline satisfactory	Borderline Unsatisfactory
Good	Poor
Excellent	Inferior

Together, the holistic rating scale and the PIRS gave three common data points which provided a basis for determining assessor and examiner agreement; specifically, there was a PIRS score, a PIRS pass-fail decision and a PIRS-based holistic rating score from each assessor and examiner for each test taker.

The same holistic rating scale was also attached to each checklist (CL-1 and CL-2) and rating scale (CRS-1, CRS-2, SRS-1 and SRS-2). These holistic rating scale data were used for determining the cut scores.

Only one skill, history taking, was sampled. History taking was chosen because it is the first and most fundamental step for most patient assessments. There is also a great deal of variability in how clinicians carry out this task, depending on personal style, clinical experience, specialty training and the nature of the patient problem. History-taking (as opposed to physical examination skill or patient management) is therefore more vulnerable to assessor bias and provides a more rigorous testing ground for scoring instruments.

The first criterion for choosing cases was the history-taking task and the choice was limited to the 12 cases in a specific administration of the OSCE. Of the 12 cases in the OSCE, history taking is a task for only three to five cases in any one administration. Fortunately, there were two cases from the fall 2003 administration that were matched not only by task but also by patient gender, medical field (psychiatry) and format (i.e., both were ten-minute cases). These two similar cases were selected to minimize case effect. In the first case, the patient problem was Depression, presented by a 24-year-old woman with weight loss and stress due to poor performance at work, secondary to a depression related to the loss of her father 13 months earlier. In the second case, the problem was post-operative Delirium, secondary to alcohol withdrawal, presented by a 58-year-old woman.

The Depression case had one task, history taking, while the Delirium case involved history taking, assessing mental status and answering an oral question regarding the patient's competency to consent to treatment. Although I considered the task for the Delirium case to be more complex, the two cases were well matched on criteria such as task and patient gender. Two thirds of the Delirium checklist items were for basic history taking and the difference between the two cases did not appear critical.

Study Instruments

The three study instruments were analytic scales with items on the same metric as the PIRS and the common holistic item. The six-point format falls within recognized measurement guidelines for rating scales and basing the new instruments on this metric provided assessors with a consistent format across the scoring instruments (Clark & Watson, 1995; Mertler, 2001; Moskal & Leydens, 2000). The items within each scale

were based on dimensions derived from the checklists for each case, the basic components of medical history-taking (Lynn, Bickley, & Szilagy, 2004) and clinical expertise theory (Schmidt et al., 1990; Schuwirth & van der Vleuten, 2004). The descriptive anchors in the research instruments were worded to focus raters on aspects of expertise such as knowledge-driven judgment, rather than on thoroughness. The score for each item ranged from zero to five and anchors were provided for the end points and the middle score to define the range of relevant behaviour for each item. The overarching purpose of each study instrument was to capture assessor judgments about the relevancy of the information being elicited by the test taker.

The development process had four steps. First the scales went through a content review by the test committee responsible for the examination. Nine members of this committee were physicians with medical school faculty positions, experience with other examining bodies and/or with post-graduate degrees in education. One member was a bioethicist with a hospital-based practice and a medical school faculty position. Committee members read through the scales individually and then discussed the merit of the items and the anchors for each instrument and the distinctions between the instruments. The session ran for approximately two hours.The recommendations from these content experts focused on clarifying the language of some of the anchors but did not involve substantive changes, nor did they recommend adding or removing any items within the scales.

Second, a pilot exercise was run with three experienced examiners. Two standardized patients, each trained to present one of the patient problems, were interviewed three times, once by each examiner. During each interview the other two

examiners observed and scored their colleague's performance. The observing examiners scored with either the skill-specific rating scale or the appropriate case-specific rating scale. Each type of rating scale was scored at least two times by at least two examiners. The three examiners were then asked to comment on each instrument.

There was a consensus among the three examiners that the rating scales were challenging for two reasons. Unlike the checklists, the rating scales did not provide much information about the patient's history and the rating scale items, although fewer in number, required more thought to score. The examiners also wanted to know more about the patient problem before scoring and more time to complete their scoring. Three changes resulted from this exercise. First, I drafted patient summaries that were then reviewed by the test committee secretary. This information was printed on the front of each score sheet as a reference for the assessors. Second, the wording on some anchors was refined for clarity and third, the need for assessors to familiarize themselves with the scoring instrument during the preparation time prior to the examination start was emphasized in the instructions they received. Allowing the assessors more time was not possible within the administrative constraints of the examination.

Third, a clinical psychologist involved in teaching reviewed the items to provide a non-medical review of the anchors for clarity and consistency. No changes resulted.

Fourth, a member of the test committee, a physician, also reviewed the scales to ensure that none of the edits made during the preceding steps had created clinically erroneous content. Based on the feedback received during the development process, all three scales were deemed to have sufficient content validity for the proposed studies.

While complete versions of all three rating scales are in the appendixes, some of the details for CRS-1 and the skill-specific rating scale are given here to illustrate their content and to demonstrate the difference between the instruments used in the two studies. CRS-1 had six rating items related to eliciting a history for depression. The items were:

1. Elicits relevant history of patient's weight loss and poor work performance,
2. Elicits factors associated with primary diagnosis of depression,
3. Assesses relevant past medical history,
4. Assesses impact of depression on patient's work and personal life,
5. Elicits history regarding father's death, and
6. Explores appropriately the patient's perception/comprehension of her problems.

The anchors were specific to each item and the qualifying terms within that item. For example, eliciting a *relevant* history of the patient's presenting problem requires the test taker to exercise clinical judgment about what he most needs to know about this patient's problem. The instructions test takers received prior to beginning the history-taking task provided sufficient information for them to be able to identify depression as the most likely primary diagnosis. They were expected to make this clinical judgment and to organize their approach to the patient accordingly. Therefore the anchors for each rating item focused the assessors on rating the test takers ability to demonstrate that they had made this judgment and were acting accordingly, including attending to key features of the problem. Establishing the severity of the depression and the risk of suicide are important for this patient and the anchors for the second item, "elicits factors associated with primary diagnosis of depression", were:

- No assessment of degree of depression (e.g., risk of suicide) and/or ignores contributing factors; e.g., medications, drugs, alcohol; poor focus.
- Partial assessment of degree of depression (e.g., risk of suicide) and/or contributing factors; e.g., medications, drugs, alcohol.
- Clear line of thought in assessing degree of depression (e.g., risk of suicide) and/or contributing factors; e.g., medications, drugs, alcohol.

The skill-specific rating scale had seven items that described critical tasks in medical history-taking, without any links to a specific patient problem. The items for this scale were:

1. Elicits relevant history of the presenting problem,
2. Elicits relevant history of associated symptoms,
3. Elicits other history relevant to the differential diagnoses,
4. Elicits a history of medications,
5. Elicits a relevant past medical history,
6. Elicits a relevant psychosocial history, and lastly,
7. Explores the patient's perception of the problem.

The anchors were specific to the qualifying terms for each item but not to any one presenting problem and therefore did not include any criteria specific to key features. The principle of focusing the assessors on test takers' ability to demonstrate an approach to the patient based on exercising clinical judgment was the same. For example, for the third item, related to diagnosis, the anchors were:

- Absent or irrelevant review of systems and history of contributing factors; no clear line of thought.
- Limited review of systems and/or history of contributing factors; line of thought may be unclear

- Relevant review of systems and history of contributing factors, shows a clear line of thought

The number of rating items for each of the rating scales was determined in large part by the nature of the history-taking task and for Study One, took into account the patient problem. Creating more items for any of the rating scales would have led to artificially splitting apart key aspects of the task and did not appear justified. Fewer items would have led to a holistic scale which would not have served the purpose of these two studies. Practically speaking more than seven or eight items would have been a challenge for the assessors as they also had to score the seven-item PIRS within a very brief amount of time.

Checklists

Checklists for each case had been developed prior to this research initiative. The development process for these instruments was similar to the one described for the study instruments.

The checklist for Depression had been used in four previous OSCE but only once before in a large-scale administration (n>500) with a cohort similar to that of the fall 2003 administration, specifically the fall 1994 administration. The checklist was modified slightly after the 1994 administration. One item regarding drugs and alcohol was changed. On the 1994 checklist drug and alcohol use was represented by one item. In 2003 this item was split into three; one each for alcohol use, prescription drug use and non-prescription drug use. In the 2003 version of the checklist, there were 20 items detailing the points of information that were deemed important to a relevant history of this patient's problem.

The checklist for Delirium had been used in five previous OSCE, three times with large cohorts (n>500) similar to that of the fall 2003 administration. The checklist had 26 items detailing the relevant information and these were the same items used in the spring 2002 OSCE.

Table 2 provides a summary for both checklists across two OSCE administrations. (As a note, the number of test takers for 2003 includes the participants from Studies One and Two, as well as test takers from an additional four sites (n=324).) The descriptive statistics indicate that each case performed similarly in 2003 to a previous administration, which confirms that the extra observer in the room (as required for this study) did not impact noticeably on test taker performance. The same table provides the ITC for the score from each checklist (as 1 of 20 checklists in 1994 and 1 of 12 after that) and these values were sufficiently high (>.20) (Nunnally & Bernstein, 1994) to indicate that each case was assessing performance that was consistent with that being assessed by the overall OSCE in both years and that the checklist scores were discriminating among the test takers (McDonald, 1999). The α-coefficients were derived from the items *within* each checklist for the study data only and suggest weak (for CL-1, with 20 items) to moderate (for CL-2, with 26 items) internal consistency, as seen in Table 2.

Table 2 *Checklist means, SD, cut score, pass rate, ITC, α and Φ*

	Year	Study	N	Mean (%)	SD (%)	Cut Score (%)	Pass Rate (%)	ITC	α	Φ
CL-1 (20 items)	1994		702	75	13	67	78	.31[a]		
	2003	One	778	76	12	67	79	.47[b]	.55	.51
		Two	741	76	11		80	.33[b]	.47	.37
CL-2 (26 items)	2002		699	61	14	55	67	.37[b]		
	2003	One	778	61	21	56	64	.41[b]	.65	.57
		Two	741	61	21		65	.39[b]	.60	.51

[a] *As one of 34 OSCE items*
[b] *As one of 23 OSCE items*

The higher α- and Φ-coefficients for CL-2 are consistent with the larger number of items. The difference in the pass rates between CL-1 and CL-2 is evidence that Delirium is a more difficult case, probably due to the difference in tasks as discussed earlier. Since the checklist scores were not being used to describe test taker performance but rather to make pass-fail decisions, internal consistency of the scores for each checklist was less important than the reliability of the decisions. Coefficients of dependability (Φ) (Brennan, 1983) were calculated for the study data based on variances taken from generalizability studies run in urGenova (Brennan, 1999) using a p x I design where p was test takers and I was items. These values indicated moderate dependability for the checklist pass-fail decisions, except for CL-1 which had a lower coefficient (Φ=.37) with the Study Two cohort. Both α- and Φ-coefficients were lower for Study Two, suggesting a possible cohort effect, even though the two cohorts were very similar based on demographic variables (e.g., age, gender, Canadian versus international training).

Patient Interaction Rating Scale (PIRS)

The PIRS was used to score behavioural aspects of the test takers' interactions with the standardized patients in six of the OSCE cases, including the two selected for this research. This instrument was previously developed and validated by the Medical Council test committee responsible for the OSCE content and has been used in conjunction with checklists to score the ten-minute patient cases where the task is to take a history for the past five years.

Both the assessors and the examiners scored the PIRS, which had seven items. Each item was a six-point rating of behaviours such as "listening skills", "organization of interview" and "rapport with person". Only the PIRS data collected for these studies were

available. However, when these data were initially assessed, one item within the scale, "closure", had not performed as well as the other items. As a result it was removed from the scale for both studies. The reliability and validity of the modified PIRS were both improved by this change. The analysis that led to this decision is reported in the Results chapter, under Item Analysis.

Assessor Feedback Form

A feedback form was provided to each assessor. The form asked assessors to provide data describing themselves: Faculty position, age, gender, discipline, previous experience with the Part II examination, and experience with other examinations (e.g., examinations of the Royal College of Physicians and Surgeons or the College of Family Physicians of Canada). Then the form asked how frequently they had used similar rating scales for other assessments and to rate the study instruments based on whether the anchors provided sufficient information, the items comprehensively captured test takers' performance, and if they had a preference for scoring a checklist or a rating scale. There was also space for the assessors to offer any general comments they wished to make. A sample of this form is in Appendix E.

Procedures for Data Collection

Data were gathered from 12 university sites across two administrations of the OSCE, all run on the same day. At each site, the OSCE was run in one or more tracks. Each track was a series of clinic rooms of sufficient number to administer the examination to one cohort of test takers per half day. Within each track, the clinic rooms were organized into two circuits. One circuit was used to administer short cases (i.e., five minutes with a patient and then five minutes completing a written task) which comprised

one half of the examination. For each of these cases there were two score sheets, one for the patient encounter and one for the written task that were added together to provide a case score. This circuit is only relevant here insofar as it explains how the capacity of 32 test takers per track per administration was achieved and why the number of items is greater than the number of cases. The other circuit contained two duplicate series of long cases (i.e., ten minutes with a patient). Depression and Delirium were both long cases. In one series, data for Study One were collected. In the other series, data for Study Two were collected. For all but one of these cases there were two score sheets, a case-specific checklist and a PIRS.

To understand how the OSCE ran, consider the path taken by one test taker, Jane Smith. She starts in Case 3 of Series A of the Long Case circuit shown in Figure 2.

In the first ten minutes, Jane completes the assigned task for Case 3, after which she moves to Case 4, to a pilot case and then to Case 5 and so on. After completing Case 7 she goes to Case 1 and Case 2. At this point she has completed one half of the examination, along with seven other test takers rotating through the same series of rooms.

Concurrently, eight other test takers complete the same set of cases in Series B. Now Jane and the fifteen test takers who have completed the Long Case circuit move to the Short Case circuit. (The sixteen test takers who completed the Short Case circuit do the reverse and now complete the Long Case circuit.) Jane again starts in Patient Case 3. The test taker from the other Long Case 3 waits. After completing Patient Case 3, Jane moves to Written Case 3 and answers written items regarding the patient she just saw. Concurrently, the Case 3 test taker who was waiting is now engaged with Patient Case 3. Next Jane moves through two rest stations and then goes to Patient Case 4, Written Case

4, and continues until she has rotated through Patient Case 2 and completed Written Case 2. She has now completed the OSCE, along with 31 other test takers.

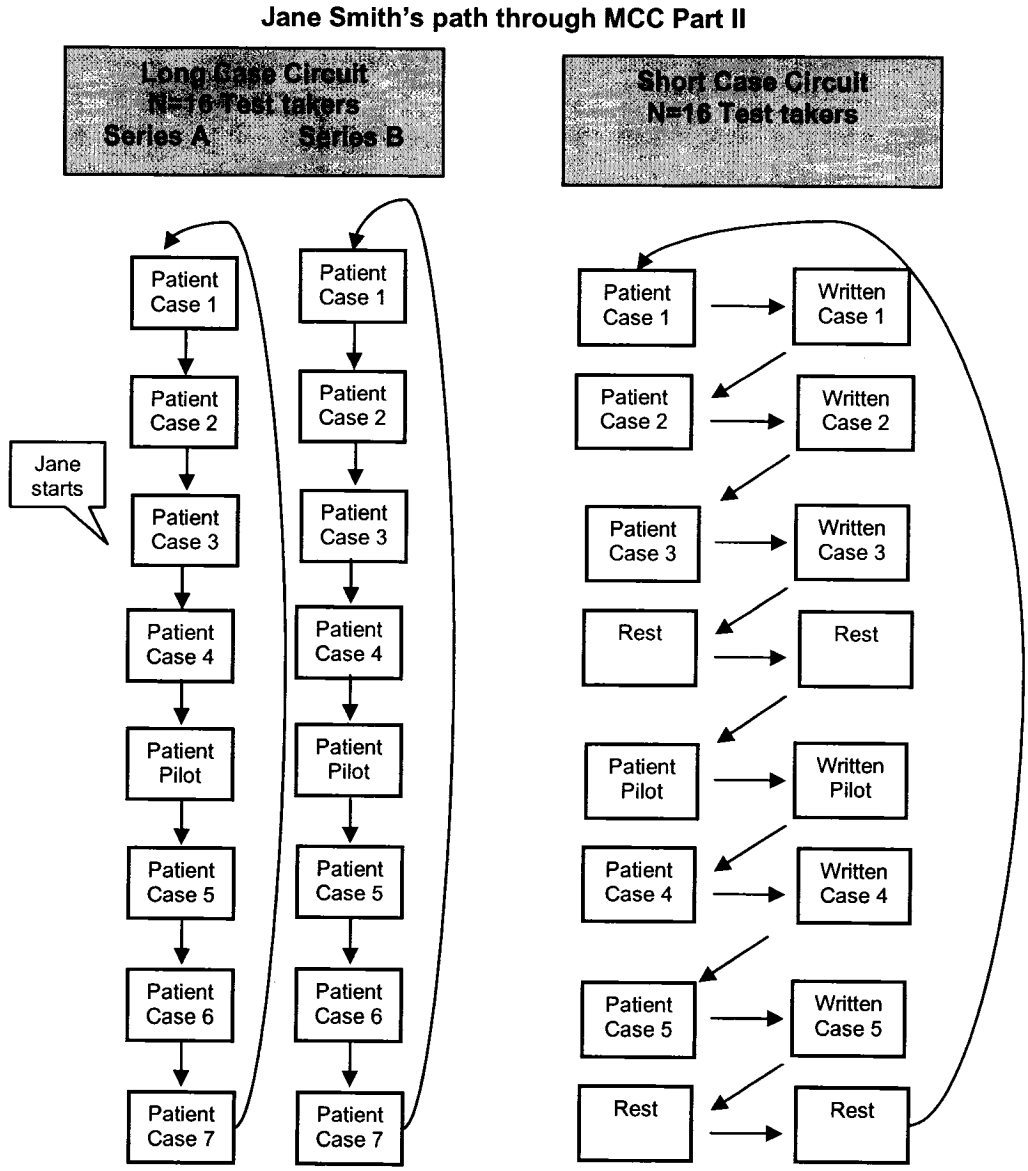


Figure 3 Test taker pathway

Table 3 is an overview of the performance data collected at *one* track during *one* administration of the OSCE and the data collected from each participant. At most sites the OSCE was administered simultaneously across two or more tracks with different

assessors for each half day. Data were collected from 27 tracks in the morning and again in the afternoon.

Guidelines for assigning physicians to cases were not rigid. In general, physicians were not assigned to cases related to their subspecialty or to cases so far from their specialty as to make it difficult for them to score accurately. Assessors scored two different cases over the course of the day. Within these two guidelines, their assignment to a case was driven more by local logistics than by any other criteria. The preference for these studies was to use experienced examiners as assessors so that they could make a meaningful comparison between the use of checklists and rating scales. Where new examiners were selected as assessors (due to local recruitment issues) they were assigned first to a non-study case so that they would have some experience with an existing checklist from a non-study case before scoring from one of the study instruments.

Just before or on the examination day, the assessors attended a locally run orientation which included instructions to review the patient problem and the scoring instruments. Instructions also included a directive to protect the integrity of the study by not conferring with the other physician in the room, who would be scoring with a checklist.

Assessors were asked to consider participating in the study prior to being assigned to a case with the intent that only those who chose to participate would be assigned to the study instruments. Consent forms (see Appendix F) were signed prior to reaching the room where they would be scoring.

Table 3 *Data collected from one track during one OSCE administration*

Case	Series	Instruments (1 set / Assessor)	n[a]	Performance Data	Test Taker Data	Assessor Data
Depression	A	CL-1 + PIRS CRS-1 + PIRS CL-1 + PIRS	16	Score per Instrument Global rating per Instrument	Knowledge – based score	Specialty Previous Part II experience
	B	SRS-1 + PIRS	16	Pass / Fail per Instrument	Clinical experience (Years since graduation) Specialty Clinical skill score	Previous experience with other exams Previous experience with rating scales Feedback on rating scales * Sufficient time? * Anchors? * Relevance of items? * Scales vs. Checklists?
Delirium	A	CL-2 + PIRS CRS-2 + PIRS CL-2 + PIRS	16	Score per Instrument Global rating per Instrument		
	B	SRS-2 + PIRS	16	Pass / Fail per Instrument		

[a]*Maximum number of test takers per administration*
Note. Test taker and Assessor data were collected per individual, not per administration

Although the local examination centres were able to recruit all of the required assessors, not all of the physicians that were required for the OSCE and this research arrived at every site on the day of the OSCE (i.e., a few forgot, were ill, or were responding to a personal emergency). In these instances, a few of the physicians recruited as assessors were used as examiners for part or all of the day, which is the major reason why the data sets have uneven numbers of test takers and assessors.

For unspecified reasons, not all of the assessors who agreed to the study signed consent forms allowing the use of their demographic data or personal observations regarding the rating scales. A few who had consented did not complete the feedback form, while some who had not consented to the use of these data did in fact provide it.

Following the examination, a follow-up mailing was sent to this latter group asking if they would consent to their data being included in the study, which did improve the participation rate to over 90% for both cases across both studies.

The actual number of test takers who registered for the OSCE was somewhat lower than the maximum and their numbers were not as evenly balanced across case and study as the description in Table 3 would suggest. A randomly uneven distribution of test takers across the two studies occurred at a few sites where the number of test takers who registered for the site was below that site's capacity. In addition, a few test takers were moved from their assigned circuit before or during the examination to accommodate local logistical problems (e.g., an ill standardized patient or temporarily absent examiner). However, these events are common to a large-scale OSCE and did not systematically impact the cohort for either study.

Despite minor irregularities on the day of the OSCE, the data were collected at the sites as planned and returned immediately afterwards for analysis.

Plan for Data Analysis

There were a number of steps that led to the comparison of scoring instruments. In total, the steps for each study were:

1. Treatment of missing data. (As this step had to be completed before any analysis could occur, the details are reported in the next section.)
2. Item analysis of the scoring instruments. (e.g., item frequencies, principle component analysis (PCA), item difficulty and generalizability studies).
3. Calculation of independent variables, cut scores for the study instruments and dependability coefficients for each cut score.
4. Control for rater effect.

5. Run regression analyses with the pass-fail decisions for each instrument as the dependent variable and expertise variables as predictors.
6. Summarize assessor feedback.

Treatment of Missing Data

A total of 1,587 test takers were assigned to cases where data for the two studies were being collected. Given the human factors involved, complete data were not collected in all instances, largely due to an assessor or examiner not fully completing a rating scale or checklist. For instance, sometimes a global rating was missing and sometimes a score for one or more items within a rating scale were missing and sometimes data from one or another scoring instrument were not collected. For each instrument, missing data were identified and were replaced or the record deleted from the study, depending on the nature and quantity of what was missing. The details of these decisions are reported by type of instrument, beginning with the checklists.

Checklists.

The checklists for both cases were scored dichotomously. Examiners filled in the bubble next to an item to indicate that the task had been "Done Satisfactorily". An empty bubble indicated the item had not been completed satisfactorily so in effect there were no missing values for these data and only the holistic ratings for these instruments were assessed for missing data. If the holistic rating was missing from either a checklist or its associated PIRS, the record for the test taker was removed as these data were critical to the planned comparisons. Twenty-eight records were removed for this reason, 16 from Depression and 12 from Delirium. Holistic rating data were also missing from two of the three rating scales and from the three PIRS. As a result of these omissions a further set of

records were deleted, although for each data set the number of records deleted per instrument ranged from one to ten (<1%).

Data from one or the other of the two cases were missing for a further 171 test takers, largely due to a shortage of examiners at three of the sites. (The shortfall meant assessors became examiners and no data for the study were collected.) Records were removed from another 88 test takers because of multiple missing data points (i.e., more than two missing) within one scoring instrument, indicating assessor or examiner error. Finally, 34 records were deleted because more than one assessor was paired with one examiner, creating very small data sets for each assessor (e.g., n<8). This occurred due to scheduling challenges at a site where successive substitute assessors scored as the staff tried to solve an internal problem. There were no patterns evident in either the missing data or the examiner mis-assignments and the deletions were deemed to be minor given the size of the data sets.

Most deletions involved removing a test taker's record from one *or* the other case. From the original cohort of 1,587 test takers, only 19 records for test takers were removed from both data sets, leaving the number of participating test takers at 1,568.

CRS-1 and CRS-2.

For Study One, data from CRS-1 and CRS-2 were based on 811 test takers. There were data from both instruments for 743 test takers. (Data for both instruments were not available for all 811 due to the missing data problems and deletions from the checklists as described above.) The comparisons for CRS-1 were based on 778 test takers and the comparisons for CRS-2 were based on 776 test takers.

The rating scale data (from CRS-1 and CRS-2) that were retained also suffered from multiple instances of missing data, but only from one or two items out of the six items within each rating scale. Leaving the missing data as missing would have significantly impacted the total score on that instrument for each of the affected test takers. Removing any record missing one or two data points would have reduced the cohort by more than 100 test takers. Therefore, the missing values were replaced with the mean score for that test taker on that instrument. By using the test taker's own mean score, the impact of replacing the missing data on their score for that case was minimized and the meaning of the score represented by the other items was unchanged. The percentage of replaced data points per instrument ranged from less than 1% up to 2% with one exception (Item 5 on CRS-1), where 4% of the 4,704 data points were missing. In total, values were imputed for 167 of 25,360 (.7%) data points.

Skill-specific rating scale.

In the case of Study Two the skill-specific rating scale data were based on 757 test takers, with data from both cases for 654. The comparisons for SRS-1 were based on 741 test takers and the comparisons for SRS-2 were based on 670 test takers.

The skill-specific rating scale used in Study Two also had multiple instances of data missing from one or two items within any one rating scale score and again, scores were imputed for those items based on the mean score for that test taker on that instrument. The percentage of replaced data points per instrument ranged from less than 1% to 1.7%. In total, values were imputed for 105 of 24,344 (.4%) data points.

Independent Variables.

Data from independent variables included test taker age and gender, year of graduation, field of training, post-graduate training group (Canadian or international), total score from the Medical Council of Canada's Part I[1] examination, and total OSCE score. Three independent variables, Knowledge, Experience and Skill, were used as a model of expertise for the regression analyses. Formal knowledge is a prerequisite for expertise and was represented by the Part I score. Experience integrated with that knowledge is critical to expertise and was represented by years since graduation from medical school. The development of relevant skills is also relevant and was represented by two sub-scores that were created for each test taker by subtracting their score for either Depression or Delirium from their total OSCE score. The hypothesis is that an increase in any one of these variables would increase the probability of passing any one station. Those with the highest values for all three would be the most likely to pass any one station. There were no missing data for age, gender, year of graduation, post-graduate training group or for total OSCE score. However, not all of the test takers reported a field of training. Field of training data were missing for 79 (9.7%) of the Study One test takers and missing more frequently for the international cohort. There were missing field of training data for 55 (7.3%) of the Study Two test takers and again, it was missing more often for the international cohort than for the Canadian trainees.

---

[1] The Part I examination is a knowledge-based test administered at the end of undergraduate medical programs. Three quarters of the content is administered as multiple-choice questions. One quarter is administered in a short-answer write-in format that assesses clinical decision-making.

Assessor Data.

Personal data and feedback regarding the rating scales were collected from the assessors. In Study One, forty-five assessors for Depression (90%) consented to their personal data and feedback being used. Of these, eight were missing data for age and one assessor did not complete the feedback form. Forty-eight of the assessors for Delirium (96%) also consented to be participants. Of these assessors, eight were also missing data for age but all had completed the feedback form. In Study Two, forty-two assessors for Depression (86%) consented to be participants. Six were missing age data but all completed the form. Thirty-nine assessors for Delirium (89%) consented to be participants although five of them did not provide age data.

Following the treatment of missing data an item analysis was conducted for each scoring instrument. The methodology for this step is described next.

Item Analysis

Evidence of content validity for all three of the scales (CRS-1, CRS-2, and the skill-specific rating scale) came from the development process, which was reported in this chapter under Instruments. Item analyses were completed after the OSCE to investigate for further evidence of validity and reliability.

For each study instrument, the frequencies for the scores (i.e., 0 to 5) for each item within a scale were reviewed to assess whether the full range of judgments had been invoked. Item difficulty was considered. Mean scores are one indicator of difficulty and with this OSCE, the mean scores have always been higher for first-time takers than repeat takers, as has the pass rate, another indicator of difficulty. Therefore, the descriptive statistics for the rating scale scores were examined across test taker groups (i.e., first time

takers versus repeater takers) to see if these scores discriminated appropriately relative to the checklist scores.

The ITC values for each of the items *within each scale* were indicators of the degree to which each item contributed to that rating scale. The general cut-off point of acceptability was an ITC>.20 (Nunnally & Bernstein, 1994). A lower value indicates that an item score correlates so poorly with the total score that its inclusion in the instrument is questionable. The α-coefficient for the rating scale score was an indicator of the instrument's internal consistency. The standard error for α and its confidence intervals were calculated as they contributed to comparisons of the alpha coefficients across instruments. The estimates of the standard error of α are sensitive to multidimensionality or covariance heterogeneity. As items within an instrument become less uniformly correlated, the standard error increases and therefore the width of the confidence interval increases (Duhachek & Iacobucci, 2004).

Principle component analyses (PCA) were run for each instrument to further assess their dimensionality and aided the interpretation of the results of subsequent analyses. PCA is a common approach to assessing whether the number of items in a set of data can be reduced to a smaller number of meaningful components and is a measure of the dimensions within an instrument. The details of the PCA are reported in Appendix G. In short, the PCA results were consistent across instrument format. All three study instruments and the PIRS were unidimensional. Both checklists were multidimensional. Running PCA for the checklists was questionable because the items were dichotomous and for some items the ratio "done" to "not done" was high (e.g., 90:10) (McDonald, 1999; Tabachnick & Fidell, 1996) and inter-item correlations were almost universally

under .30 (Tabachnick & Fidell, 1996). These limitations lower the value of interpreting the components, but not the usefulness of PCA for determining the dimensionality of the scores (McDonald, 1999). The implications of the differences in dimensionality are considered with the results.

Generalizability studies were run using a p x I design where p was test takers and I was items. Within this model, the proportion of variance attributed to test takers indicates how well the scores associated with each instrument are representing test taker performance. The studies were run in urGenova (Brennan, 1999). The variance components from these studies are reported as part of the comparison of the distribution of scores (e.g., variance ratios, degree of skew) across instrument and case.

Independent Variables and Cut Scores

After the item analyses, the next step was to generate the Experience variable (i.e., years since graduation from medical school), followed by the calculation of the two Skill scores, and the determination of the cut scores, which led to the dependent pass-fail variable for each instrument. Last was the calculation of the dependability coefficients. A brief description of the procedures for all but the calculation of Experience follows.

Skill scores.

One of the independent variables selected as a marker of expertise was a skill-based score, which was derived from the OSCE scores. The Medical Council's OSCE is a validated performance assessment of the important clinical skills that all physicians are expected to possess (Dauphinée et al., 2000; Poldre et al., 1999; Reznick et al., 1997; Reznick, Blackmore, Dauphinée, Rothman, & Smee, 1996). In order to use scores from this OSCE as a predictor variable, two skill scores were calculated for each test taker, one

for each case which excluded that case. Therefore, the skill score for Depression was the sum of the scores for the other eleven cases (including Delirium). The skill score for Delirium was the sum of the scores for the other eleven cases (including Depression).

Cut scores and dependability coefficients.

In addition to the Experience and Skill variables, cut scores for the study instruments had to be calculated so that the dependent variables could be generated. This was accomplished using the same borderline group method as was used for the checklists scores. With this approach, the assessors were the standard setters. When assessors classified test takers as "Borderline Satisfactory" or "Borderline Unsatisfactory" on the holistic rating scale, they identified performances that defined the pass standard. Their judgments were translated into a cut score by taking the mean score of the borderline test takers, which became the passing standard for each instrument.

Once cut scores were established, the dependability of the pass-fail rating scale decisions could be assessed and compared to the dependability of the checklist pass-fail decisions. For this task, coefficients of dependability (Φ) (Brennan, 1983) were calculated for each cut score using variances from the generalizability studies.

The calculation of the Φ-coefficients completed the initial data analysis. The data were cleaned, the scores from each instrument had been assessed, the independent variables were all in place, and the dependent variables had been generated. The next step was to control for rater effect through the selection of test taker records for the regression analyses. How the criteria for these decisions were arrived at is described next.

Control for Rater Agreement

The purpose of both studies was to ascertain whether there was an instrument effect, which meant that rater effect had to be minimized. To this end only test taker records where rater agreement existed were included in the regression analyses. For each test taker, the examiner and assessor each scored one common instrument, the PIRS. By only selecting records for test takers where the two raters agreed on the PIRS, it was more likely that the results of each comparison of checklist scores to rating scale scores could be attributed to an instrument effect rather than to rater differences.

Several indices of inter-rater reliability were considered. Correlations based on the PIRS scores, the global ratings and the pass-fail decisions, using Pearson's *r*, Spearman's rho and Cohen's kappa respectively, were all significant. However, the correlations were all less than .5. In some ways, finding only moderate correlations with these measures was not surprising given that test takers at this level of education and training (i.e., minimum of 12 months of post-graduate clinical training) are expected to perform well and most do. The effect is a negatively skewed distribution which attenuates the upper limit of correlation that can be observed (Stemler, 2004). Kappa values indicated less than moderate agreement (i.e., range of .157-.254 for scores and .269-.397 for pass-fail decisions) based on a correction for a 50:50 probability of agreement by chance. However, with negatively skewed pass-fail data the probability of agreeing by chance is also skewed, making it difficult to interpret a kappa coefficient. The results of the correlation analyses and the Kappa values are summarized in Appendix H for reference but in the end, none of these indices were used as an index of inter-rater reliability.

After considering the above measures, an alternate approach to controlling for assessor effect was taken. Instead of using a correlation, two other criteria were used to identify consensus between the examiners and the assessors. For each PIRS score there was also the holistic rating running from Excellent (1) to Inferior (6). Therefore the first criterion for rater agreement was agreement on the holistic rating item. For the purposes of this study, agreement was defined as ratings within one point of each other. So long as a difference in rankings was not greater than one point, the paired examiners and assessors were deemed to agree. Based on this criterion, differences could not be greater than the difference between poor and borderline unsatisfactory or borderline satisfactory and good. This criterion excluded records where an examiner and assessor difference as great as that between excellent and borderline existed. Second, the pass-fail outcome for the PIRS had to be the same. The double criteria represented the summative nature of the assessment and the focus on pass-fail decisions. In this context, consensus on decisions was emphasized over consistency of scores (Stemler, 2004).

Using these criteria, cohorts of test takers were selected for each case within each study based on rater agreement and the regression analyses run on their outcomes, which was the next step in the plan.

Regression Analyses

For this research, a model of expertise comprised of three independent or predictor variables representing critical components of expertise was used to compare the pass-fail decisions across instruments. The three variables were Knowledge, Experience and Skill. The strength of the relationship between the predictors of expertise and the pass-fail decisions was the basis for comparing the study instruments to the checklists. If

the model more accurately predicted the pass-fail decisions for one instrument, it would suggest an instrument effect and would be evidence of greater discrimination for that instrument based on its sensitivity to expertise and would support a construct validity argument for the use of that instrument format. If there were no difference in the accuracy of the pass-fail predictions between instruments that would be evidence in support of a utility-based validity argument favouring whichever instrument was easiest to develop and score. The next section explains why logistic regression was employed and how the results are interpreted.

Rationale for logistic regression.

For each instrument, the dependent variable was the pass-fail decision, which is a dichotomous outcome. Regression models are used to predict the mean of the dependent variable, Y. When Y is dichotomous, the model is predicting the *probability* that a case will fall into the category coded as one or in these studies, fall into the category of Pass. While the *predicted probability* of Pass can be infinitely small or large, the observed probability of Pass occurring changes very little as the predictor variable(s) reach more extreme values. Stated another way, as the probability of Pass occurring approaches zero or one, changes in the value of the predictors have less and less impact. This is a nonlinear function, better described by an S-shaped line.

A logistic transformation is one of the simplest ways to represent an S-shaped line and is achieved by taking the natural logarithm of the ratio of the probability that Y=Pass to the probability that Y Pass (1- Pass) (Menard, 2001; Pampel, 2000). The ratio of Pass:(1-Pass) is the odds that Y=Pass and the logistic transformation of the odds is the logit or logged odds of Y=Pass. With the logit function, the resulting variable ranges

from negative infinity (as the odds of Y=Pass decrease) to positive infinity (as the odds of Y=Pass increase). Estimated probabilities do not exceed zero or one. The logit function is similar in appearance to a linear regression function and the interpretation of the logit function is also similar. For example, a one unit increase in any one of the predictor variables multiples the odds of Y=Pass by the odds ratio for that variable, which is the coefficient (β) for that variable, exponentiated.

The exponents of the coefficients (Exp (β)) or odds ratios are reported as part of the results of the logistic analyses and are used to compare the contribution that each predictor variable makes to the model. When an odds ratios is greater than one then the odds of Y=Pass increase as the predictor variable increases. When an odds ratio is less than one then the odds of Y=Pass decrease as the predictor variable increases (Menard, 2001).

In these studies the underlying assumption of the model was that increases in a predictor (indicating more knowledge, skill or experience), would increase the odds of passing. If the three-variable model was more accurate at predicting decisions for one or the other instrument, that would be evidence of greater discriminant validity for that instrument.

The interpretation of the logistic regression analyses was based on indicators and statistics which are described briefly in Appendix I. A brief guide to interpreting these statistics is given here:

1. Coefficient (β) for each predictor
2. Standard error of measure for β (SE)
3. Wald statistic for each predictor

4. Significance of Wald statistic
5. Odds ratio for each β, shown as Exp(β).

The sign (positive or negative) of the coefficient is one indicator of the direction of the change in the odds of passing given a one unit change in the predictor and the Wald statistic is a two-tailed test of significance for each coefficient. The odds ratio indicates the degree of change in the odds of passing, given a one unit change in the predictor. For example, if Exp(β)=1.135 for Skill with the CL-1 pass-fail decision as the dependent variable then a one score increase in the Skill score increases the odds of passing CL-1 by 13.5%. By considering these measures in conjunction with one another, conclusions about the relationship between the model and the dependent variable can be made and contribute to the comparisons of the scoring instruments.

Summarizing Assessor Feedback

Some of the assessor feedback was demographic and was reported in the Methodology chapter under Participants. As well, assessors were asked to rate five items (on a scale of one to five where one was lowest and five was the highest rating):

1. Time allowed for scoring,
2. Previous use of similar scoring instruments,
3. Quality of the anchors,
4. Comprehensiveness of the scoring instrument, and
5. Preference for rating scale relative to checklist.

Lastly, assessors were asked for any general observations about the rating scales. The scores were summarized in histograms (Appendixes K and L) and the comments reviewed for common themes. Highlights of the findings are reported with the results and were included in the comparisons, particularly across studies. Summarizing the assessor feedback was the final step in the overall analysis. The results follow.

Results

The results are reported in chronological sequence as each step depended on the previous one. There are five major components to this chapter.

1. Item analysis of each scoring instrument, beginning with the PIRS. Included in the item analysis are descriptive statistics, ITC, α-coefficients and the results of the generalizability studies.
2. Logistic regression analyses, four for each study, to support comparisons of the pass-fail decisions. The results are reported by study. Reported with the regression results are correlations of rating scale scores with the relevant checklist scores and the correlations for the pass-fail decisions. The correlations were explored as criterion measures of validity because the checklists were the standard against which the rating scales were being assessed. An external criterion would have been a stronger indicator of concurrent validity but no such criterion was available.
3. Alternate models of expertise were considered and the results of these are reported because two of the predictor variables contributed little to the original model.
4. Comparisons were made across the two studies using results that are summarized in Table 27.
5. A summary of the assessor feedback concludes the chapter.

Item Analysis

Patient Interaction Rating Scale

The PIRS is shown in Appendix B. As indicated earlier, the fifth item, Closure, was removed from the PIRS. The decision was triggered by the comparatively low ITC for this item across all test taker cohorts (i.e., study by case), as shown in Table 4.

Table 4 *ITC values for Closure item compared to ITC values for other PIRS items*

Case	Study	Marker	Closure ITC	Min ITC for Other Items	Max ITC for Other Items
Depression	One	Examiner	.48	.67	.81
		Assessor	.49	.62	.79
	Two	Examiner	.54	.62	.81
		Assessor	.50	.56	.80
Delirium	One	Examiner	.46	.50	.72
		Assessor	.54	.55	.78
	Two	Examiner	.47	.54	.74
		Assessor	.51	.54	.74

The pattern of ITC values was considered in conjunction with comments from ten of the assessors that strongly suggested the time limit faced by the test takers made closing the interview appropriately a problem. To quote one assessor, "Closure of the interview had to be abrupt .... the interview ended at the 9 minute mark so no one was able to end the interview properly." Given that the time limit artificially ended at least some interviews, the validity of the Closure item was questionable. Some test takers never actually demonstrated the behaviour and others would have done so badly due to the time limit, yet all the test takers had a score for this item. Given that this behaviour was poorly sampled, the Closure scores were unlikely to accurately reflect test taker ability and excluding Closure was justified on the grounds of content validity alone. As removal of this item also improved the internal consistency of the scores for the PIRS, the shortened version was used for both studies.

Descriptive statistics and α-coefficients for the PIRS minus Closure are presented in Table 5. The ITC for PIRS as one of 23 items *within the OSCE* are all moderately high. Of note, within this OSCE there were a total of six PIRS used, including the two for this research. For Depression, the scores from the other five PIRS comprised 13.7% of the

total OSCE score and for Delirium, the scores from the other five cases made up 14.5% of the total OSCE score. The ITC values for items *within PIRS* were high and α-coefficients for the instrument as scored by assessors and examiners across two cases were also high (>.80), indicating strong internal consistency for the six items that were retained.

Table 5 *Mean score, SD, ITC, α-coefficient & correlations for PIRS*

	Depression		Delirium	
	Examiner	Assessor	Examiner	Assessor
		Study One		
N	778		776	
Mean (%)	80.3	78.3	71.7	72.7
SD (%)	13.3	14.3	14.7	16.0
Variance Ratio (%)	16.6	18.3	20.5	22.0
ITC (PIRS within OSCE)	.42	.49	.46	.51
α (OSCE; N_{Items}=23)	.81	.82	.80	.80
ITC Min (Items within PIRS)	.70	.65	.53	.57
ITC Max (Items within PIRS)	.83	.80	.75	.80
α (PIRS; n=6)	.92	.91	.84	.87
Correlation[a]	.46	.82	.54	.70
	(CL-1)	(CRS-1)	(CL-2)	(CRS-2)
		Study Two		
N	741		670	
Mean (%)	79.0	79.7	76.3	72.7
SD (%)	15.3	15.0	14.3	15.3
Variance Ratio (%)	19.4	18.8	18.7	21.0
ITC (PIRS within OSCE)	.47	.49	.38	.45
α (OSCE; N_{Items}=23)	.81	.81	.81	.81
ITC Min (Items within PIRS)	.65	.57	.57	.59
ITC Max (Items within PIRS)	.82	.82	.77	.77
α (PIRS, n=6)	.92	.91	.86	.87
Correlation[a]	.40	.74	.46	.68
	(CL-1)	(SRS)	(CL-2)	(SRS)

[a]*Pearson correlation with checklist and ratings scores were all significant (p<.01, two-tailed)*

The Pearson correlations for the PIRS scores to the checklist scores and the study instrument scores are also reported in this table. These correlations were one basis for comparing scores from the checklists to scores from the study instruments. Of note, the

PIRS scores were more correlated with the study instrument scores than with the checklist scores, as seen when the correlation for the examiner-scored PIRS is compared to the assessor-scored PIRS. The difference is more pronounced for Depression, suggesting a case effect, as seen in Table 5. The variance components as calculated from generalizability studies run in urGenova (Brennan, 1999) are reported in Table 6 and also provide evidence of a case effect. For Delirium, the variance components associated with test takers were smaller than for Depression across both sets of raters. However, the variance components for both item and the item-test taker interaction for Delirium are greater than for Depression, indicating a larger proportion of error variance in scores for Delirium regardless of who scored the test taker. Understanding the case effect is important because it can be a confounding factor when comparing scores for an instrument effect.

Table 6 *Variance components (%) for PIRS*

Case	Rater	σ^2 (tt)%	σ^2 (i)%	σ^2 (tti)%	Total %
Depression	Examiner	62.3	5.6	32.1	100
	Assessor	58.6	8.3	33.1	100
Delirium	Examiner	42.8	13.7	43.5	100
	Assessor	46.4	11.2	42.4	100

Note. tt = test taker, i = item and tti= test taker – item interaction

The dimensionality of the PIRS was assessed by running a PCA of the PIRS scores for each case within each study. In each instance only one component was extracted, based on eigenvalues >1, with 60-61% of variance being explained by a first component solution.

The combination of α-coefficients, assessor comments and PCA results were taken as evidence that the shortened version of this rating scale produced scores that were sufficiently content valid and internally consistent enough to be used in the research.

Case-specific Rating Scale-1

The CRS-1 is shown in Appendix C. The results of the item analysis for CRS-1 are presented in Table 7, along with the results for CL-1.

Table 7 *Mean, SD, ITC, pass rates, α and correlations CL-1 & CRS-1*

	CL-1	CRS-1
# of Items	20	6
Mean (%)	72.8	75.2
SD (%)	12.5	16.0
ITC Min (Items within instrument)	.071	.610
ITC Max (Items within instrument)	.308	.760
ITC (Instrument within OSCE[a])	.334	.477
Mean (%) FTT / RT*	73 / 68	76 / 65
Effect size for difference Cohen's d[e]	.46	.71
Pass Rate (%) FTT / RT[b]	82 / 66	85 / 62
Effect size for difference Cohen's d[e]	.41	.62
α	.554	.866
SE (α)	.023	.007
95% CI for α[c]	.509 – .599	.852 - .880
Score correlation (CL-1 to CRS-1)[d]	.299	
Pass-Fail correlation (CL-1 to CRS-1)[d]	.232	

[a]*For CL-1, α=.809, for CRS-1, α=.817; N_{Items}=23 for both*
[b]*FTT (First time takers): n=707, RT (Repeat takers): n=71.*
[c]*Confidence Intervals (Duhachek & Iacobucci, 2004)*
[d]*Pearson correlations, all significant at p<.01, two-tailed*
[e]*Small Effect (>=.15 and <.40); Medium effect (>=.40 and<.75); Large effect >=.75)(Thalheimer & Cook, 2002)*

All six items from CRS-1 were retained. Frequency studies showed that the full range of judgments within each item *within the instrument* were used and scores for all items were negatively skewed, which was the expected distribution for test takers at this level of training (i.e., all test takers had successfully completed a minimum of 12 months

of post-graduate training). The mean score for CRS-1 fell within the expected range of difficulty for the ten-minute cases across all test takers (N=1,911), which was 63% to 77%. The ITC values for the items *within the instrument* were also high, well above .20, the general cut-off point of acceptability. When CRS-1 was treated as an *item within the OSCE* (with scores from CL-1 removed) then the ITC for CRS-1 scores was an indicator of the degree to which CRS-1 assessed a facet of the same construct as the OSCE (α=.817, n=23). For CRS-1, ITC = .477, which was well above the minimum standard and was higher than ITC=.334 for CL-1. (As a note, to isolate the ITC values for CRS-1 and CL-1 from their respective PIRS scores, the analyses were done on 23 component scores, as all but one of the 12 case scores was comprised of scores from two different instruments.)

The mean score for first-time takers was significantly higher than for repeat takers for CL-1 and CRS-1 ($p<.002$, two-tailed *t*-tests), as was expected. The mean pass rate was also significantly higher for first time takers for both instruments ($\chi^2 = 9.884$, $df=1$, $p<.002$ and $\chi^2 = 22.868$, $df=1$, $p<.000$ for CL-1 and CRS-1 respectively). The difference between these two groups was greater for CRS-1 than for CL-1 (see Cohen's d), which indicated greater discrimination.

The α-coefficient for CRS-1 was significantly higher than for CL-1, based on the lack of overlap between their respective confidence intervals. In addition, the confidence interval for CRS-1 was narrower than for CL-1, indicating stronger inter-item correlations. Although not shown, the inter-item correlations for CRS-1 fell between .42 and .72 (mean =.53), reflecting a moderate degree of correlation between each item and the five other items and indicating that the estimate of the α-coefficient was relatively

accurate (Duhachek & Iacobucci, 2004). Although the Pearson correlation between CL-1 and CRS-1 scores was significant, it was not high. Similarly the correlation between CL-1 and CRS-1 pass-fail decisions was significant but not high. These correlations hint at the same conclusion that Reznick and his colleagues (1998) reached; namely that the two instruments are measuring different constructs or they are measuring the same construct with a marked degree of error.

Taken as a whole, the item analysis provided further evidence of validity for CRS-1 (e.g., level of difficulty, discrimination between first time takers and repeat takers) and strong evidence of internal consistency (α=.866).

Case-specific Rating Scale-2

The CRS-2 is shown in Appendix C. A summary of the item analysis for CRS-2 is in Table 8, along with the results for CL-2. All six items from CRS-2 were retained. Frequency studies showed that the full range of judgments *within each item* within the instrument was used and scores for all items were negatively skewed, which was the expected distribution for these test takers. The mean for CRS-2 fell within the expected range of difficulty for this examination (i.e., 63% to 77%, as reported above), although the mean score for CRS-2 was higher than for CL-2. As with CRS-1, the ITC values for the items *within the instrument* were high.

When CRS-2 was treated as an item *within the examination* (with scores from CL-2 removed) then the ITC for CRS-2 scores was a measure of the degree to which CRS-2 assessed a facet of the same construct as the OSCE (α=.804, n=23). For CRS-2, ITC = .390 was well above the minimum standard and it was modestly higher than ITC=.335 for CL-2. (As noted earlier, to isolate the ITC values for CRS-2 and CL-2

from their respective PIRS scores, the analyses were done on 23 component scores, as all but 1 of the 12 case scores are comprised of scores from two different instruments.)

Table 8 *Mean, SD, ITC, pass rates, α and correlations for CL-2 & CRS-2*

	CL-2	CRS-2
# of Items	26	6
Mean (%)	56.5	66.4
SD (%)	14.2	19.0
ITC Min (Items within instrument)	.070	.359
ITC Max (Items within instrument)	.352	.656
ITC (Instrument within OSCE[a])	.335	.390
Mean (%) FTT / RT[b]	57 / 54	67 / 60
Effect size for difference Cohen's d[c]	.21	.39
Pass Rate (%) FTT / RT[b]	61 / 55	71 / 48
Effect size for difference Cohen's d[c]	.12	.50
α	.644	.792
SE (α)	.018	.011
95% CI for α[d]	.609 - .679	.770 - .814
Score correlation (CL-2 to CRS-2)[e]	.500	
Pass/fail correlation (CL-2 to CRS-2)[e]	.341	

[a]*For CL-2, α=.809, for CRS-2, α=.804; N_{Items}=23 for both*
[b]*FTT (First time takers): n=710, RT (Repeat takers): n=66*
[c]*Small Effect (>=.15 and <.40); Medium effect (>=.40 and<.75); Large effect >=.75)(Thalheimer & Cook, 2002)*
[d]*Confidence Intervals (Duhachek & Iacobucci, 2004)*
[e]*Pearson correlations all significant at p<.01, two-tailed*

CRS-2 discriminated between first time takers and repeat takers, as indicated by the significant difference in the mean scores for the two groups (p<.002, two-tailed t-test) and a significant difference in the mean pass rates (χ^2 = 14.96, df=1, p<.000). Although the means and pass rates for the two groups were similarly different for CL-2, the difference in pass rate was not significant (χ^2 = 1.1, df=1, p<.295). As with CRS-1, this was evidence that CRS-2 discriminated better than did CL-2 and is corroborated by Cohen's d statistic.

The α-coefficient for CRS-2 was significantly higher than for CL-2, based on the lack of overlap between their respective confidence intervals and as with CRS-1, the confidence interval for CRS-2 was narrower than for CL-2. The inter-item correlations (not shown) for CRS-2 were examined and fell between .19 and .59 (mean=.40). As for CRS-1, these values indicated that each item had a modest degree of correlation with each of the five other items, indicating that the calculation of the α-coefficient was reasonably accurate. The correlation between CL-2 and CRS-2 scores was significant and sufficiently high to suggest that the two instruments were assessing facets of the same construct. The correlation between the pass-fail decisions for the two instruments was also significant but not as high.

As with CRS-1, the level of difficulty, discrimination between first-time takers and repeat takers was additional evidence of validity for CRS-2 and was matched by strong evidence of internal consistency (α=.792). The similarity in the results across the two cases is good evidence that the case-specific rating scale scores are more valid and reliable than the checklist scores, even without any further analysis. Regardless, as part of the process of comparing these instruments to the checklists, generalizability studies were run using a basic model of person by items where persons were the test takers. The variance components are reported in Table 9.

Table 9 *Variance components (%) from Study One*

Case	Instrument	σ^2(tt)	σ^2(i)	σ^2(tti)	Total %
Depression	CL-1	4.4	24.7	70.9	100
	CRS-1	49.9	3.9	46.2	100
Delirium	CL-2	5.3	19.3	75.4	100
	CRS-2	37.5	3.4	59.1	100

Note: tt = test taker, i = item, and tti = test taker – item interaction

This table shows a difference in test taker variance components between the rating scale scores and the checklist scores although the difference is not as large for Delirium, which suggests a case effect. The variance components for item, which represents error associated with the instrument, are larger for the checklists than for the rating scales, indicating an instrument effect.

Histograms, variance ratios and the degree of skew for each instrument are shown in Figure 4 and show that the differences between the two types of instrument lie at both extremes of the score distributions. The vertical reference lines indicate the cut score.

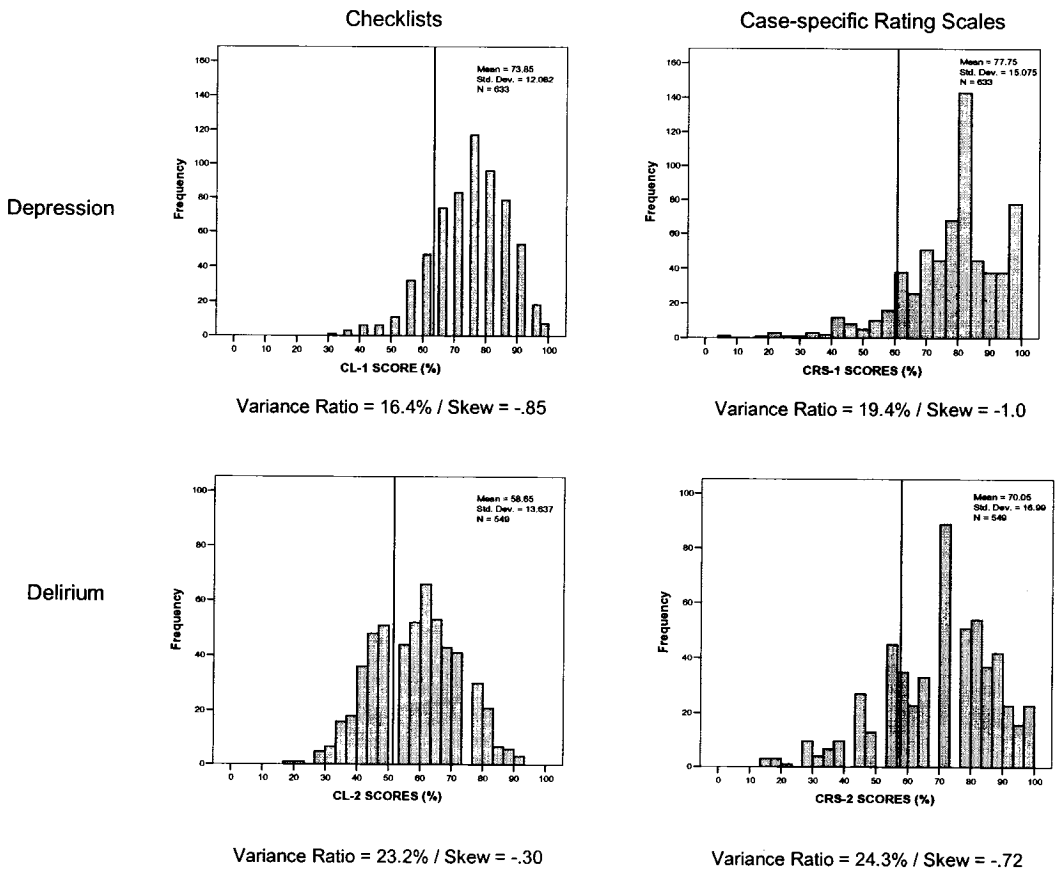


Figure 4 Score distributions for Study One

For example, for CRS-1, shown on the upper right of the figure, there are more test takers in both tails of the distribution than there are in the distribution for CL-1, shown in the upper left of the figure. For Depression and Delirium, the rating scale scores have a wider range, a larger variance ratio, and are more skewed than the checklist scores. (Please note that the gaps in the histograms are an artefact of using percent scores rather than raw scores.)

The pattern of differences between the checklist scores and the case-specific rating scores holds across case (e.g., the rating scale scores for both cases have higher internal consistency and a higher variance ratio), demonstrating an instrument effect that is evident despite the case effect (indicated by the difference in means scores, cut scores and pass rates, and variance ratios). Further evidence of an instrument effect is the greater dependability of the pass-fail decisions made with the study instruments. In addition, the study instrument scores discriminated as well as or better than the checklist scores, as shown by the mean score differences between first time takers and repeat takers.

Skill-specific Rating Scale

The seven skill-specific rating scale items (see Appendix D) were aimed at history-taking skill without reference to a patient problem. As a result, the same skill-specific rating scale was used for both cases and the item analyses of the skill-specific rating scale for Depression and Delirium are summarized together in Table 10. All seven items from the skill-specific rating scale were retained.

Frequency studies showed that the full range of judgments *within each item* within the skill-specific rating scale was used, although all the scores for all the items were negatively skewed. The means for the skill-specific rating scale were close to the

expected range of difficulty for this examination (the same values given under CRS-1, namely 63% to 77%), although the mean for SRS-2 was below the lower limit. The ITC values for items *within the instrument* were all strong as indicated by the minimum and maximum values. When the skill-specific rating scale was treated as an item *within the OSCE* (with the relevant checklist removed) then the ITC for a skill-specific rating scale score was a measure of the degree to which the rating scale assessed a facet of the same construct as the OSCE (α=.816 for SRS-1 and α=.815 for SRS-2, N_{Items}=23).

Table 10 *Mean, SD, pass rates, α-coefficient, and correlations for SRS*

	Depression		Delirium	
	CL-1	SRS-1	CL-2	SRS-2
# of Items	20	7	26	7
Mean (%)	72.3	65.2	55.8	55.2
SD (%)	11.5	12.4	13.5	15.7
ITC Min (Items within instrument)	.005	.466	-.036	.579
ITC Max (Items within instrument)	.264	.657	.397	.675
ITC (Instrument within OSCE[a])	.409	.454	.379	.367
Mean (%) FTT/RT[b]	73 / 68	66 / 60	56 / 52	56 / 50
Effect size for difference Cohen's d[c]	.45	.48	.29	.35
Pass Rate (%) FTT/RT[b]	81 / 68	86 / 77	64 / 47	75 / 63
Effect size for difference Cohen's d[c]	.33	.25	.35	.27
α	.470	.837	.599	.856
SE (α)	.028	.009	.022	.008
95% CI for α – Lower Limit[d]	.415	.903	.556	.840
95% CI for α – Upper Limit[d]	.525	.938	.642	.872
Score Correlation (CL-1 or CL-2 to SRS)[e]	.356		.411	
Pass/fail Correlation (CL-1 or CL-2 to SRS)[e]	.269		.315	

[a] *α=.814, α=.816, and α=.815 for the checklists, SRS-1 and SRS-2 respectively*
[b] *Depression: FTT (First time takers): n=668, RT (Repeat takers): n=73*
Delirium: FTT (First time takers): n=608, RT (Repeat takers): n=62
[c] *Small Effect (>=.15 and <.40); Medium effect (>=.40 and<.75); Large effect >=.75)(Thalheimer & Cook, 2002)*
[d] *Confidence Intervals (Duhachek & Iacobucci, 2004)*
[e] *Pearson correlations, all significant at p<.01, two-tailed*

As noted with Study One, isolating the ITC values for the skill-specific rating scale scores and the checklist scores from their respective PIRS scores led to scale analyses being done on 23 scores, as all but 1 of 12 case scores are comprised of scores

from two different instruments. For SRS-1, ITC=.454 and for SRS-2, ITC=.367, both values above the minimum acceptable value of .2.

For Depression, the mean score for first-time takers was significantly higher than the score for repeat takers for both the checklist and SRS-1 ($p<.000$, two-tailed t-test). The first time takers' pass rates were also higher, but not quite so significantly ($\chi^2 =6.74$, $df=1$, $p<.009$ and $\chi^2 =4.75$, $df=1$, $p<.029$ for CL-1 and the SRS-1 respectively). For Delirium the pattern repeated with mean scores being significantly higher for first time takers for both instruments ($p<.033$ and $p<.009$, two-tailed t-test for CL-2 and the SRS-2 respectively). The mean pass rate was also significantly higher for first time takers for both instruments ($\chi^2 =7.39$, $df=1$, $p<.007$ and $\chi^2 =4.01$, $df=1$, $p<.045$ for CL-2 and for SRS-2 respectively). These data are summarized in Table 10.

The α-coefficients for the skill-specific rating scale scores were high for both cases and higher than the values for the scores from either checklist. Although the Study Two coefficients for CL-1 scores ($\alpha=.470$) and CL-2 scores ($\alpha=.599$) were lower than the coefficients for the scores from the *same* instruments in Study One ($\alpha=.554$ and $\alpha=.644$), the differences fell within overlapping confidence intervals at the 95% level and were not deemed to be significant. The mean inter-item correlations (not shown) were .45 (Depression) and .47 (Delirium), which indicated a moderate degree of correlation among items. The correlations between CL-1 and CL-2 scores and the skill-specific rating scale scores were significant but not high, which was also true for the correlation of the pass-fail decisions.

In summary, the item analysis provided further evidence of validity for the skill-specific rating scale scores based on level of difficulty, discrimination between first time takers and repeat takers and included strong evidence of internal consistency. The weak correlations between the skill-specific rating scale scores and the checklist scores and between the skill-specific rating scale and checklist pass-fail decisions suggested that the there was only a weak relationship between what each of the two instruments is measuring, either due to measurement of different constructs or to error.

The results for the skill-specific rating scale scores were compared to checklist results across two cases. To augment the comparison generalizability studies were run using a basic model of person by items where persons were the test takers. The variance components are reported in Table 11.

Table 11 *Variance components from Study Two generalizability*

Case	Instrument	σ^2(tt)	σ^2(i)	σ^2(tti)	Total %
Depression	CL-1	3.1	27.2	69.7	100
	SRS	39.9	5.7	54.4	100
Delirium	CL-2	4.3	21.2	74.5	100
	SRS	44.0	4.4	51.6	100

Note: tt = test takers, i = items, and tti = test taker – item interaction

The pattern seen in Study One was repeated with variance components for the test takers being markedly higher for the skill-specific rating scale scores than for the checklists scores. Likewise the variance components for item were larger for the checklist scores than for the skill-specific rating scale scores, indicating that more error variance can be attributed to the instrument for the checklist scores than for the rating scale scores.

The percentage score distributions, variance ratios and degree of skew for instruments across case are shown in Figure 5. Although the variance ratios did not differ

greatly between the instruments for each case, the ratios for the skill-specific rating scale scores were slightly higher than the checklist score ratios and the skill-specific rating scale score distributions were more negatively skewed than the checklist score distributions, which was a muted repetition of the pattern of distributions found in Study One. The vertical reference lines are set at the cut score for each instrument. (Please note that gaps in the histograms are an artefact of reporting in percent scores rather than in raw scores.)

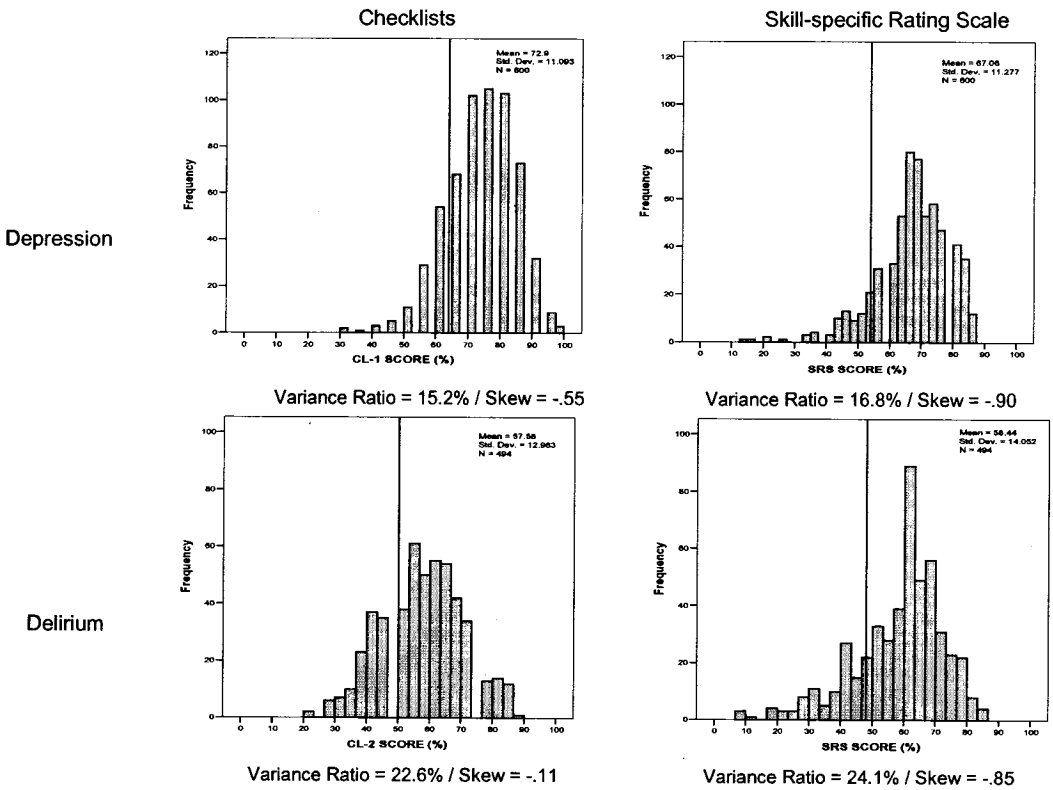


Figure 5 *Score distributions for Study Two*

The pattern of differences between the checklists and the skill-specific rating scale shown in Figure 5 was similar to the pattern seen in Study One but the differences between the two types of instruments were less in Study Two. For example, the negative

skew was greater for the skill-specific rating scale but the difference in the variance ratios was not as great as observed in Figure 4 for Study One.

Despite being less obvious, the pattern is further evidence that there was both an instrument effect and a case effect. The larger variance ratios for the skill-specific rating scale scores, combined with a higher degree of internal consistency and more reliable decisions than occurred for either CL-1 or CL-2, suggests an instrument effect. The lower mean score for Depression, regardless of the instrument, indicates that Depression was less difficult than Delirium, which is a case effect.

Cut Scores and Pass-Fail Decisions

The cut scores and pass rates for the checklists and study instruments for Study One were set using the borderline group method as described earlier and the outcomes are reported in Table 12. Within a case there is not much difference in the number of test takers identified as borderline or in the pass results. There is however a notable difference across case. The borderline groups were larger for Delirium and pass rates were lower, even with a lower cut score, all of which indicated Delirium was a more difficult case.

Table 12 *Cut scores, pass rates and Φ-coefficients for checklists & study instruments*

	Study One				Study Two			
	Depression[a]		Delirium[b]		Depression[c]		Delirium[d]	
	CL-1	CRS-1	CL-2	CRS-2	CL-1	SRS-1	CL-2	SRS-2
# of items	20	6	26	6	20	7	26	7
Borderline N	210	232	350	324	180	214	310	303
Cut Score (%)	63.5	60.7	51.5	57.7	63.9	53.8	50.4	48.1
Pass Rate (%)	80.3	82.5	60.6	69.5	80.0	85.3	62.7	73.6
Φ-coefficient	.507	.858	.570	.785	.373	.945	.574	.871

[a]*Test Taker N = 778; Assessor N = 50* [c]*Test Taker N = 741; Assessor N = 49*
[b]*Test Taker N = 776; Assessor N = 50* [d]*Test Taker N = 685; Assessor N = 44*

As shown in Table 12, the Φ-coefficients for the checklist cut scores were consistently lower than for the study instrument cut scores. The differences were more pronounced for Study Two, suggesting that an instrument difference between the case-specific rating scales and the skill-specific rating scales exists.

Regression Analysis

Within each study, analyses were run to assess the accuracy of an expertise model with three independent variables for predicting pass-fail decisions for each instrument. Comparisons of the model's accuracy across instruments were made to assess the relative validity of the study instruments within the context of this model. If the model fit and it more accurately predicted pass-fail decisions for one type of instrument over another, this would be evidence for a validity argument for that instrument because it would indicate that its pass-fail decisions represented better discrimination within the expertise construct. As described in the Methods chapter, logistic regression was used for this step in the research because the dependent variable (i.e., pass-fail decision) was dichotomous. The results are reported by study.

Prior to conducting the regression analyses, the data were checked for outliers. Univariate outliers were defined as independent variables with an absolute z-score >= 3. Multivariate outliers were defined as independent variables with a Mahalanobis distance greater than 16.266 ($df = 3$; $p<.001$) (Tabachnick & Fidell, 1996). All analyses were run using SPSS 13.0 (2004).

Study One

In Study One, the question was whether the case-specific rating scale pass-fail decisions were more accurate or valid than the checklist pass-fail decisions. One way to

make the comparison was to see if an expertise model, in this case represented by the three independent variables of Knowledge, Skill and Experience, better predicted the outcomes for one or the other type of instrument. Since pairs of raters were involved in scoring, differences between the scores could be attributed to either instrument differences or rater differences. To limit the risk of rater differences contributing to (or masking) instrument differences, only those records where there was inter-rater agreement on the PIRS were included in the regression analyses. If raters agreed on the common scale then differences in the scores from the checklists and the rating scales were more likely due to instrument differences. For the Depression case, 633 records (81% of total records) were retained while for the Delirium case, 549 records (71% of total records) were retained. The difference in cohort size between the two cases indicates that there was less rater agreement with Delirium and was another sign of a case effect.

Once the test taker records had been selected, the data were checked for outliers. For Depression, only one record was identified as an outlier, due to a high Knowledge score ($z = 3.05$). For Delirium, five records were identified as outliers; two due to low skill scores ($z=-3.0$), and three due to a high number of years since graduation (Experience: $z=3.0$, 3.3 and 4.3). All records were checked for accuracy and retained.

CRS-1.

Goodness of fit is the first concern with any model because it indicates whether predictions are improved by the coefficients for the independent variables, which were estimated from the data set. If the fitted model is significantly more accurate in predicting the dependent variable than would occur by chance (i.e., if the coefficients were zero), then the model can be used to explain and compare the dependent variable. Two

measures of fit are given in Table 13. The full model χ^2 compares the model's predictions to chance predictions whereas the Hosmer and Lemeshow measure compares the model's predictions to an idealized model. Both show that there was a significant degree of model fit even though R_L^2 showed that the model only explained a small percentage of the variance, meaning the association between the predictor variables and the dependent variable was not strong. Given that limitation, the values for R_L^2 still showed a stronger association between the model and CRS-1 pass-fail decisions than between the model and the CL-1 pass-fail decisions. (See Table 13.)

Table 13 *Goodness of fit and measures of association for CL-1 & CRS-1*

	P/F for CL-1			P/F for CRS-1		
n=633	**Statistic**	***df***	**Sig**	**Statistic**	***df***	**Sig**
Full Model χ^2	32.59	3	.000	78.19	3	.000
Hosmer & Lemeshow	4.49	8	.810	2.07	8	.979
Nagelkerke R^2 / R_L^2	.084 / .127			.220 / .164		

Note. P/F = Pass-Fail

After considering goodness of fit, the predictions for the dependent variables were compared to the observed values. Table 14 shows that the total accuracy of the predictions was almost the same for the two instruments and the τ_p values indicated that the results were significantly accurate for both. However, the model better predicted fails for CRS-1 than for CL-1 (i.e., 12.7% correct versus 1.9% correct for Fails).

Table 14 *Observed to predicted Pass-Fail decisions for CL-1 & CRS-1*

Instrument		**Fail Pred**	**Pass Pred**	**% Corr**	**Total % Corr**	**τ_p**	**Sig (*d*)**
CL-1	**Fail Obs**	2	104	1.9	83.4	.41	6.52
	Pass Obs	1	526	99.8			
CRS-1	**Fail Obs**	10	69	12.7	87.8	.50	7.15
	Pass Obs	8	546	98.6			

Note. Obs=Observed, Pred = Predicted and Corr = Correct

Despite the pattern of significance and accuracy for the model, only Skill showed evidence of being a significant contributor, as seen in Table 15. Skill appears to have more power to predict CRS-1 decisions as indicated by Exp (β) = 1.126 but the overlapping confidence intervals for Skill indicated this was not a significant difference.

Table 15 *Comparing Knowledge, Experience and Skill for CL-1 and CRS-1*

DV = Pass-Fail Decision for CL-1								
Predictors	**β**	**SE**	**Wald**	***df***	**Sig**	**Exp(β)**	**95% CI Exp (β) (Lower**	**Upper)**
Knowledge	.002	.002	1.433	1	.231	1.002	.999	1.006
Experience	-.009	.018	.238	1	.625	.991	.958	1.026
Skill	.067	.021	1.041	1	.002	1.069	1.026	1.115
DV = Pass-Fail Decision for CRS-1								
Knowledge	.003	.002	1.297	1	.255	1.003	.998	1.007
Experience	-.029	.019	2.372	1	.123	.971	.936	1.008
Skill	.119	.025	22.836	1	.000	1.126	1.072	1.182

Note. β = coefficient for variable; SE=standard error of β; Wald is test of significance for β and Exp(β) is the odds ratio for β.

CRS-2.

The results of the logistic regression analyses for Delirium began with the goodness of fit indicators given in Table 16. The χ^2 statistic for the model was significant for CL-2 and CRS-2, while the Hosmer and Lemeshow statistic showed no significant difference between the model for either instrument and an "idealized" model. The model explained little variance for either instrument but the R_L^2 statistic indicated that the model explained more variance for CRS-2 decisions than for CL-2 (e.g., 16.4% versus 5.7%).

Table 16 *Goodness of fit indicators and measures of association for CL-2 & CRS-2*

	P/F Decision for CL-2			P/F Decision for CRS-2		
n=549	**Statistic**	***df***	**Sig**	**Statistic**	***df***	**Sig**
Full Model χ^2	4.83	3	.000	47.08	3	.000
Hosmer & Lemeshow	3.24	8	.918	3.02	8	.933
Nagelkerke R^2 / R_L^2	.100 / .057			.125 / .164		

Note. P/F = Pass/Fail

Table 17 shows the model predicted decisions for CL-2 and CRS-2 with much the same degree of accuracy, as indicated by the τ_p statistic. While the model better predicted decisions for CRS-2, the overall pattern was the reverse of that seen with Depression. With Delirium, the model better predicted checklist fails and study instrument passes.

Table 17 *Observed to predicted Pass/Fail decisions for CL-2 & CRS-2*

Instrument		Fail Pred	Pass Pred	% Corr	Total % Corr	τ_p	Sig (*d*)
CL-2	**Fail Obs**	33	150	18.0	68.9	.41	1.04
	Pass Obs	21	345	94.3			
CRS-2	**Fail Obs**	14	109	11.4	78.0	.44	8.18
	Pass Obs	12	414	97.2			

Note. Obs=Observed, Pred = Predicted and Corr = Correct

Table 18 shows that only Skill was a significant predictor for both instruments. Experience was significant for the checklist and indicated that a year's increase in Experience improved the odds of passing by 4.2%. Although the model had a better fit and was more accurate for CRS-2, Skill appeared to have more power to predict checklist pass-fail decisions (i.e., a higher Exp (β) value). However, this finding was not significant given the overlapping confidence intervals.

Table 18 *Knowledge, Experience and Skill for CL-2 & CRS-2*

Predictors	β	SE	Wald	*df*	Sig	Exp(β)	95% CI Exp (β) (Lower	Upper)
DV = Pass-Fail Decision for CL-2								
Knowledge	-.001	.001	.222	1	.638	.999	.997	1.002
Experience	.041	.018	5.266	1	.022	1.042	1.006	1.080
Skill	.114	.021	3.699	1	.000	1.121	1.076	1.167
DV = Pass-Fail Decision for CRS-2								
Knowledge	.000	.002	.013	1	.909	1.000	.997	1.003
Experience	-.012	.018	.413	1	.520	.988	.954	1.024
Skill	.099	.022	2.290	1	.000	1.104	1.058	1.153

Note. β = coefficient for variable; SE=standard error of β; Wald is test of significance for β and Exp(β) is the odds ratio for β.

In the regression analyses for Study One, Knowledge was the one predictor variable that was never significant. The second predictor, Experience, was almost never significant and the third, Skill, was always significant. However, the differences between the Exp (β) values for Skill across instruments were not significant. With the easier case, Depression, the study instrument produced more internally consistent scores with a more dependable cut score and the relationship between Skill and the pass-fail decisions was stronger than for the checklists. For Delirium, the results are very similar. The study instrument produced more internally consistent scores with a more dependable cut score and the model better represented the pass-fail decisions. The similarity of results across the two cases supports the conclusion of an instrument effect but the differences between the cases suggest there was also a case effect.

Study One Correlations.

The nature of both instrument and case effect were further explored by considering the correlations among the scores from each of the instruments, as summarized in Table 19. As the degree of measurement error for each instrument varied

(indicated by the α-coefficients on the diagonal), observed correlations (reported in Appendix J) were corrected for attenuation (Crocker & Algina, 1986 (pp. 236-238)).

Table 19 *Score correlations corrected for attenuation across instruments for two cases*

	CRS-1	CRS-1 PIRS	CL-1	CL-1 PIRS
CRS-1	.87			
CRS-1 PIRS	**.923**[a]	.91		
CL-1	**.432**	.355	.55	
CL-1 PIRS	.471	**.522**	**.652**[a]	.92
	CRS-2	**CRS-2 PIRS**	**CL-2**	**CL-2 PIRS**
CRS-2	.79			
CRS-2 PIRS	**.841**[a]	.87		
CL-2	**.703**	.428	.64	
CL-2 PIRS	.452	**.502**	**.730**[a]	.84

[a]*Over-correction accounts for correlation being higher than alpha coefficient*
Note. Pearson correlations, all significant (p < .01; two-tailed)

Given the aspects of performance that each instrument was intended to assess, there were correlations between instruments that were expected to be high or at least to be higher than other correlations because they were convergent measures. Correlations expected to indicate convergence are underlined and in bold in Table 19. Other correlations were expected to be low, or at least to be lower than the underlined correlations because these correlated scores represented more discriminant aspects of performance. Correlations representing discriminant scores are simply in bold. The most discriminant correlations are in plain font and had the lowest values, as hypothesized. However, overall the pattern for the corrected correlations did not correspond to expected patterns.

Consider the correlations between the two PIRS scores for each case. These are convergent measures and correlations should be high because the same instrument was used to assess the same domain or trait. Despite the convergence of method and domain,

these score correlations (one for each case) are only moderately strong (r=.522 and r=.502). More significantly, they are not higher than the discriminant correlations between the checklist or case-specific rating scale and its respective PIRS. Instead, the correlations between scores for each checklist or case-specific rating scale and its respective PIRS are higher than the correlation between the two PIRS scores.

The checklist and the case-specific rating scale were intended to represent a different aspect of performance than the PIRS scores, yet the two highest correlations are between CRS-1 and the CRS-1 PIRS (r=.923) and CRS-2 and CRS-2 PIRS (r=.841). These correlations are between scores generated by the same rater, from two instruments with the same format (rating scale) but whose purpose is to assess different but related dimensions of the same construct. The case-specific rating scales are supposed to capture judgments about the nature and relevance of the information that test takers gather; the key-feature dimension. The PIRS is supposed to capture judgments about the skill or behaviours that are demonstrated by the test takers as they gather that information. These were not independent dimensions but correlations as high as seen with Depression, raise the question of whether the scores from these two instruments are representing the same dimension of the domain. Even somewhat lower correlations, such as were found for CL-1 and CL-1 PIRS (r=.652) and CL-2 and CL-2 PIRS (r=.730), would have better fit discriminant expectations.

Of note, the convergent correlations between the checklist scores and the case-specific rating scale scores were higher for Delirium (r=.703), the more difficult case, than for the Depression case (r=.432). However, the checklist scores for Delirium were also more internally consistent and more strongly correlated with the PIRS scores than

was true for the Depression case (i.e., r=.730 versus *r*=.652). The case difficulty is largely due to test takers having to complete three related tasks; take a history, conduct a focused mental status examination and determine competency to consent. The complexity of the task resulted in more variance for the Delirium case, as seen in Figures 4 and 5 and this likely aided raters in distinguishing between test takers and perhaps between the task-related dimension and the patient interaction dimension. If so, then the difference in case difficulty explains the difference in the pattern of correlations between the two cases.

In summary, Table 19 indicates that scores from the study instruments are strongly enough correlated with scores from their respective PIRS to suggest that the case-specific rating scale instruments may have been more of an extension to the PIRS instrument than a replacement for the checklists. This result was more evident for the Depression case than for the Delirium case. Measurement error, indicated by lower α-coefficients for the Delirium rating scale and its PIRS, may have moderated this correlation.

Study Two

In Study Two, the question was whether the skill-specific rating scale pass-fail decisions were more accurate or valid than the checklist pass-fail decisions. Skill-specific rating scale items were generic to the skill of history taking, rather than being specific to the task within a case. Although the data for Study Two were collected from the same cases during the same OSCE as reported in Study One, the data were from a different cohort of test takers and a different cohort of examiners and assessors. Following the approach taken in Study One, comparisons across instruments were made of the predictive accuracy of the expertise model as represented by Knowledge, Skill and

Experience. To limit the risk of rater differences contributing to (or masking) instrument differences, only those records where there was inter-rater agreement on the PIRS were included in the regression analyses. If raters agreed on the common scale then differences in the scores from the checklists and the rating scales were more likely due to instrument differences.

After accounting for missing data and rater agreement, 600 test taker records (81% of all records) were retained for Depression; while 494 records (74% of all records) were retained for Delirium. For Depression, three records were identified as outliers due to high Knowledge scores (z = 3.1, 3.6, 4.3). For Delirium, one record was identified as an outlier due to the number of years since graduation (z=4.3). The outlying records were checked for accuracy and retained.

SRS-1.

The results of the logistic regression analyses of the pass-fail decisions of CL-1 and the SRS-1 begin in Table 20, which shows the degree to which the model fit the data. The χ^2 statistic was significant for both instruments, suggesting good fit. The Hosmer and Lemeshow statistic indicated no significant difference between the observed and ideal models for either instrument, further indicating model fit.

Table 20 *Goodness of fit indicators and measures of association for CL-1 & SRS-1*

	P/F Decision for CL-1			P/F Decision for SRS-1		
N=600	**Statistic**	***df***	**Sig**	**Statistic**	***df***	**Sig**
Full Model χ^2	6.30	3	.000	49.11	3	.000
Hosmer & Lemeshow	1.30	8	.245	13.16	8	.106
Nagelkerke R^2 / R_L^2	.158 / .108			.166 / . 127		

Note: P/F = Pass / Fail

However, the model explained relatively little variance for either instrument (R_L^2=10.8% for CL-1 and R_L^2=12.7% for SRS-1), suggesting only a weak association between the model and either instrument. Although the difference between the instruments was small (1.9%), the model fit for SRS-1 was slightly better than for CL-1.

As seen in Table 21, τ_p shows that the predictions were almost identically accurate but checklist fails and study instrument passes were better predicted.

Table 21 *Observed to predicted Pass-Fail decisions for CL-1 & SRS-1*

Instrument		Fail Pred	Pass Pred	% Corr	Total % Corr	τ_p	Sig (*d*)
CL-1	Fail Obs	14	91	13.3	83.2	.48	8.09
	Pass Obs	10	485	98.0			
SRS-1	Fail Obs	2	57	3.4	90.2	.47	5.41
	Pass Obs	2	539	99.6			

Note. Obs=Observed, Pred = Predicted, Obs = Observed and Corr = Correct

Table 22 shows Knowledge and Skill were contributing to the predictions of the CL-1 pass-fail decisions while Experience and Skill were contributing to the predictions for the SRS-1 pass-fail decisions. However, Knowledge and Experience do not contribute much to the predictions.

Table 22 *Knowledge, Experience and Skill for CL-1 & SRS-1*

DV = Pass-Fail Decision for CL-1								
Predictors	B	SE	Wald	*df*	Sig	Exp(β)	95% CI Exp (β) (Lower	Upper)
Knowledge	-.004	.002	5.645	1	.018	.996	.993	.999
Experience	-.008	.018	.221	1	.638	.992	.958	1.027
Skill	.139	.022	38.394	1	.000	1.149	1.100	1.200
DV = Pass-Fail Decision for SRS-1								
Knowledge	-.002	.002	.879	1	.348	.998	.994	1.002
Experience	-.056	.020	7.826	1	.005	.945	.909	.983
Skill	.105	.026	16.706	1	.000	1.111	1.056	1.169

Note. β = coefficient for variable; SE=standard error of β; Wald is test of significance for β and Exp(β) is the odds ratio for β.

SRS-2.

The model did fit the data for both CL-2 and SRS-2, judging by the significance of the χ^2 values seen in Table 23. However, the R_L^2 values again indicated only a weak association between the predictors and the decisions for both instruments. With CL-2, the model only explained 9.1% of the variance and with SRS-2 only 11.8% was explained, also shown in Table 23.

Table 23 *Goodness of fit indicators and measures of association for CL-2 & SRS-2*

	P/F Decision for CL-2			P/F Decision for SRS-2		
N=494	**Statistic**	***df***	**Sig**	**Statistic**	***df***	**Sig**
Full Model χ^2	56.44	3	.000	46.57	3	.000
Hosmer & Lemeshow	3.23	8	.919	9.1	8	.332
Nagelkerke R^2 / R_L^2	.151 / .091			.147 / .118		

Note. P/F = Pass / Fail

Table 24 reports the accuracy of the predictions for both instruments as being the same, based on the τ_p values. However, as with Depression, the model better predicted fails for the checklist decisions and better predicted passes for the study instrument.

Table 24 *Observed to predicted Pass-Fail decisions for CL-2 & SRS-2*

Instrument		**Fail Pred**	**Pass Pred**	**% Corr**	**Total % Corr**	**τ_p**	**Sig(*d*)**
CL-2	Fail Obs	41	117	25.9	70.9	.46	11.21
	Pass Obs	27	309	92.0			
SRS-2	Fail Obs	10	80	11.1	82.4	.46	7.16
	Pass Obs	7	397	98.3			

Note. Pred = Predicted, Obs = Observed and Corr = Correct

Table 25 shows that with Delirium, only Skill was a significant predictor and there was little difference in its power to predict pass-fail decisions for either instrument.

Table 25 *Knowledge, Experience and Skill between CL-2 & SRS-2*

Predictors	B	SE	Wald	*df*	Sig	Exp(β)	95% CI Exp (β) Lower / Upper	
DV = Pass-Fail Decision for CL-2								
Knowledge	.000	.001	.032	1	.859	1.000	.997	1.003
Experience	.025	.017	2.133	1	.144	1.025	.991	1.060
Skill	.127	.023	31.491	1	.000	1.135	1.086	1.187
DV = Pass-Fail Decision for SRS-2								
Knowledge	.002	.002	1.239	1	.266	1.002	.998	1.006
Experience	.024	.020	1.5115	1	.218	1.025	.986	1.065
Skill	.118	.025	21.725	1	.000	1.125	1.071	1.182

Note. β = coefficient for variable; SE=standard error of β; Wald is test of significance for β and Exp(β) is the odds ratio for β.

The regression analyses for Study Two, as with Study One, showed that Knowledge was only a significant predictor once, as was true for Experience. Skill, as before, was always a significant predictor. However, the differences between the Exp (β) values for Skill across instruments were not significant. With the easier case, Depression, the study instrument produced more internally consistent scores with a more dependable cut score but the expertise model did not offer any further evidence of construct validity.

For Delirium, the results were very similar. The study instrument produced more internally consistent scores with a more dependable cut score but the expertise model did not offer any evidence of greater validity. The similarity of the results of the item analyses across the two cases is evidence of an instrument effect, even without instrument differences being identified with the regression analyses. Similarly to Study One, the item analyses also provided evidence of a case effect, as the mean scores and pass rates for Depression (Table 10) indicate that it was an easier case than Delirium.

Study Two Correlations.

The correlations among the different instrument scores were compared to enrich the analyses already completed. As explained in Study One, correlations were expected to be higher or lower, especially relative to other correlations, depending on whether or not the performance dimensions that instruments were supposed to assess were more or less convergent. Correlations expected to indicate convergence are again underlined and in bold in the table; correlations expected to reflect discriminant dimensions of the construct are simply in bold.

Table 26 shows the correlations among the Study Two instruments. As the degree of measurement error for the scores from each instrument varied (as shown by the α-coefficients on the diagonal), the observed correlations (reported in Appendix J) were corrected for attenuation.

Table 26 *Score correlations corrected for attenuation across instruments for two cases*

	SRS-1	SRS-1 PIRS	CL-1	CL-1 PIRS
SRS -1	.84			
SRS-1 PIRS	**.846**[a]	.91		
CL-1	**.567**	.411	.47	
CL-1 PIRS	.497	**.604**	**.601** [a]	.92
	SRS-2	**SRS-2 PIRS**	**CL-2**	**CL-2PIRS**
SRS-2	.86			
SRS-2 PIRS	**.790**	.87		
CL-2	**.572**	.457	.60	
CL-2 PIRS	.505	**.502**	**.635** [a]	.86

[a]*Over-correction accounts for correlation being higher than alpha coefficient*
Note. Pearson correlations, all significant (p < .01, (two-tailed)

Table 26 shows that the skill-specific rating scale scores correlated better with their respective PIRS scores than the checklist scores did with their respective PIRS scores. Table 26 also shows that the strength of the skill-specific rating scale correlation

to its PIRS score was somewhat stronger for Depression (r=.846), the easier case, which was similar to the pattern for Study One as seen in Table 19. However, the correlation between the checklist scores and the skill-specific rating scale scores was almost the same across the two cases (r=.567 and r=.572), unlike Study One where this correlation was stronger for Delirium. In summary, the correlation between the skill-specific rating scale and the PIRS may mean that the SRS is acting more like an extension of the PIRS than like a replacement for either checklist.

One of the limitations of the regression analyses was the general lack of significance for two of the predictor variables. Before considering the results further, alternate approaches to modeling expertise were considered and are briefly discussed next.

Alternate Expertise Models

Two variables, Knowledge and Experience, did not contribute significantly to the expertise model used in the regression analyses. Knowledge was based on scores from the Part I, the only common examination of medical knowledge available for these test takers. However, test takers took this examination at different times in their training/career path and at different times relative to when they attempted the OSCE. Test takers are first eligible to take the Part I when they complete medical school and Canadian medical students almost universally complete this examination in their last term. However, international post-graduate trainees take this examination as they enter the Canadian licensure process, which may be early in their post-graduate training or after a varying number of years of clinical practice in another jurisdiction. The length of time between passing the Part I and attempting the OSCE also varies far more for

international post-graduate trainees; from less than 12 months to several years. Most Canadian post-graduate trainees attempt the OSCE within 18 months to three years of passing the Part I.

Knowledge recall changes as physicians organize their knowledge into various schemas (Schmidt et al., 1990) and as they age (Eva, 2002). While formal knowledge is an essential building block to expertise, recall of knowledge is not a marker for it (Schmidt et al., 1990). The differences in when test takers took the examination relative to their training paths and differences in age may explain why the "Knowledge" variable is not a strong predictor as these two factors would have contributed to knowledge score variance. Although there were no other measures that could be substituted, running the regression analyses on subsets of the test taker cohorts was explored. One set of regression analyses was done with test takers where the gap between graduation from medical school and completing the Part I examination was less than 10 years. Knowledge did not contribute significantly to any of these predictions. A second set of regression analyses was done with only Canadian post-graduate trainees. None of the three predictors were significant for Depression and only Skill was significant for Delirium. Appendix K has a more detailed description of these analyses but as neither approach was an improvement over the original one, no changes were made to the research.

Experience was simply the number of years between graduation from medical school and taking the OSCE and like Knowledge, it was a weak predictor. This may be because experience in and of itself does not lead to expertise; while essential it is not sufficient. Deliberate and reflective practice is key (Ericsson & Charness, 1994) and years of experience is not a direct measure of such practice. Aside from whether the

participants in this group were engaged in practice that would lead to increasing expertise, for some of them Experience also included varying periods of interrupted practice due to issues like immigration and parental leave. Furthermore, the Experience variable was an indirect measure of the role of aging in performance. With age, physicians are more likely to have an abrupt style and/or to be less up-to-date (Eva, 2002). Lack of related training, being out-of-date regarding diagnosis and treatment, and a tendency to be abrupt, perhaps exacerbated by cultural biases around gender and/or mental health problems (singly or all together) would impact negatively on a test taker's ability to competently elicit a history within the context of the two cases used in this research. These factors would explain, for instance, why the regression coefficient was sometimes negative or not significantly different from zero.

In an attempt to better represent the role of Experience in test takers' performance an alternate approach to modeling this variable was taken. Test takers were grouped according to their specialty relative to each patient problem. This approach to representing relevant experience was based on the assumption that test takers from specialties where the presenting patient problem is seen more frequently are more likely to have expertise relative to that clinical challenge. However, this approach did not change the results in any meaningful way. Skill remained the only significant predictor. More details for this alternative are also provided in Appendix K.

Although investigating alternate models based on selected test taker cohorts and specialty-based categories was an interesting exercise, these models did not perform better. The Discussion chapter is based on results from the original cohorts and model.

Comparing Across Studies

Several patterns that emerged as the data were analysed are seen in the summary for the study instruments shown in Table 27. The pattern of similar differences between cases is perhaps the most obvious one. Consistent differences for mean score, cut score and pass rate, regardless of instrument, indicate that Delirium was the more difficult case.

Table 27 *Summary of measures by study across instruments and cases*

	Study One				Study Two			
	CRS-1	**CL-1**	**CRS-2**	**CL-2**	**SRS-1**	**CL-1**	**SRS-2**	**CL-2**
	Depression		**Delirium**		**Depression**		**Delirium**	
# of Items	6	20	6	26	7	20	7	26
Mean Score (%)	75	73	66	57	79	72	65	56
Variance Ratio (%)	19	16	24	23	17	15	24	23
α-coefficient	.87	.55	.79	.64	.84	.47	.86	.60
ITC (Instrument within OSCE)	.48	.33	.39	.34	.45	.41	.37	.38
Variance explained by PCA (%)	60	53	50	64	53	51	55	61
Variance component (*tt*[a]) (%)	50	4.4	38	4.3	40	3.1	44	4.3
Score correlation (to PIRS)	.82	.46	.70	.54	.74	.40	.68	.46
Score correlation	.30		.50		.36		.41	
Pass-Fail correlation	.23		.34		.27		.32	
Cut Score (%)	61	64	58	52	54	64	48	50
Pass Rate (%)	83	80	70	61	85	80	74	63
Φ-coefficient	.86	.51	.79	.57	.95	.37	.87	.57
Φ-coefficient *of Skill Score*	.83		.83		.83		.84	
R_L^2	16.4	12.7	16.4	5.7	12.7	10.8	11.8	9.1
Odds Ratio (Skill)	12.6	6.9	10.4	12.1	11.1	14.9	12.5	13.5

[a]*tt=test taker*

The second pattern was indicative of an instrument effect. Over both studies the internal consistency of the study instrument scores was high (i.e., high α-coefficients), as was the dependability of their cut scores (i.e., high Φ-coefficients) and these measures were consistently higher than the same measures for the checklist scores. Third, the

regression model best fit the CRS-1 and CRS-2 scores, even though Skill was the only significant predictor variable across all eight analyses.

Assessor Feedback

Study One

Most of the assessors for Depression (92%) and for Delirium (96%) completed most of the feedback form. A quick summary of their responses is given in Table 28 which shows the percentage of assessors who rated the descriptor with the highest frequency. (Bar graphs further detailing these scores are in Appendix L.) Although the scores for Depression were negatively skewed, indicating support for the CRS-1, the most frequent scores were not the highest ones, and the responses seemed more reflective of tolerance than of endorsement. The scores for the more difficult case, Delirium, were less skewed with more missing data and more mid-range scores, suggesting a less supportive response to CRS-2.

Table 28 *Assessors from Study One who rated an item on its middle value*

Case	Time Sufficient	Content of Anchors Sufficient	Captured Performance Acceptably	No Preference
Depression	63%	65%	56%	33%
Delirium	52%	52%	40%	27%

For Depression, fifteen assessors added explanatory comments relevant to the study. (Other comments were about catering, parking, length of day, candidate behaviour, etc.) Two assessors expressed concern about needing more time between test takers as most of the marking was done at the end of the encounter. Comments from those preferring the rating scale included "better captures/reflects performance", "this format allowed me to pay more close attention to the interview while it was in progress",

"...More ability to separate levels with this system vs. current one, but more difficult to examine overall." Other comments were critical, for example "I find this more difficult to score than the checklist", "rating scale seems more subjective", "more difficult to assess candidates with rating scales (e.g., in terms of knowledge base)" and "checklist would be more useful to acknowledge physician's acquisition of history related to symptoms – rating scale didn't allow for accurate assessment of this."

For Delirium, sixteen assessors added relevant explanatory comments. Two of them commented on problems scoring the PIRS item for closure due to the artificial ending of the test taker-standardized patient interaction imposed by the time limit. This was the item which was removed from the analyses. Of those who commented on their preference, seven preferred the checklist. Their feedback included "Rating scales add more subjectivity to this test", "Much more demanding and difficult to use rating scale than checklist," and "Define anchors for us neophytes – much prefer checklist – it almost requires waiting to the very end to complete the rating scales." Those preferring the rating scale indicated "checklist doesn't allow you to assess skill in getting information" and "The checklist does not always reflect the candidate's ability/performance. Some items on checklist seem inappropriate / irrelevant and affect score unreasonably." The Delirium scores and comments were more critical of the case-specific rating scale than was true for Depression, suggesting case-related issues. One assessor indicated that the "instructions consistently confused candidates".

Study Two

Most of the assessors for Depression (86%) and for Delirium (89%) completed most of the feedback form. A quick summary of the responses is given in Table 29 which

shows the percentage of assessors who rated the descriptor with the highest frequency. (Bar graphs further detailing these scores are in Appendix M.) The responses for Depression were less positive than for Study One, suggesting examiners were less tolerant of the skill-specific rating scale than of the case-specific rating scale. Certainly, the scores did not indicate endorsement for the skill-specific rating scale.

Table 29 *Assessors from Study Two who rated an item on its middle value*

Case	Time Sufficient	Content of Anchors Sufficient	Captured Performance Acceptably (Partially)	No Preference
Depression	69%	57%	54%	48%
Delirium	39%	41%	(41%)	30%

Although lower, these scores were similar to Study One in that they were less positive overall for Delirium. There was also more missing data due to incomplete feedback forms for Delirium.

For Depression, fourteen assessors added explanatory comments relevant to the study. Two assessors in this study also expressed concern about needing more time between test takers to complete scoring as most of it was done at the end of each encounter. None of the comments indicated a preference for the rating scale although one assessor said it had the advantage of flexibility. He saw this advantage as offset however by the quality of the anchors. "...if you don't know medical problem at a professional experience level, there is too little information to rate performance" and by the time problem, "the candidate can't be scored until the end (therefore the examiner must recall)." Another assessor wrote

> It depends on the subject / topic being examined. For this particular station, I do not think that a generic rating scale gives an objective score. A more specific rating scale may be more useful. A generic scale like this

> one is not given to be objective when rating psychiatry or stations where the subject material doesn't necessarily follow these headings. It also takes much longer to mark. Therefore, the blocks can only be coloured in right at the end. Better communication scale - more specific.

For Delirium, eleven assessors added relevant explanatory comments. Three of these assessors reiterated the need for more time at the end of each interaction since scoring could only be completed at the end. None of the comments spoke to there being an advantage to the skill-specific rating scale. Two of the comments were critical of the case for being complex or the time being too short for test takers to complete their tasks. The last comments introduce the discussion: "...the rating scale should be less generic and more customized to the content of the station" and "Rating scale seemed open-ended, vague, makes it difficult to have standardized marking between all candidates from beginning to end."

As with the score-based data, the assessor perceptions reflect both an instrument and case effect. Common perception across studies and instruments was that the rating scales were more subjective and more time-consuming to score. The more generic rating scale, the SRS, was also described as more subjective than the case-specific rating scales.

The assessor feedback provided a different perspective on the rating scale outcomes, one that leads away from simple answers as to which instrument discriminates the best and leads directly into the discussion that follows.

Summary of Results

The results are complex and so highlights from the five sections are summarized here as a reference. Table 27 also serves as a summary for most of the findings.

1. Item Analysis

1.1. The PIRS was internally consistent (range of α was .84 to .92) and discriminating (i.e., range of ITC was .38-.51), as reported in Table 5.

1.2. The two case-specific rating scales performed similarly. Both were internally consistent with α=.87 and .79 for CRS-1 and CRS-2 respectively; ITC=.48 and .39 respectively was evidence of discrimination. These values were higher than for the two checklists (in Table 27). Further, the test taker variance components (also in Table 27) were larger for the rating scales than for the checklists. Figure 4 shows (and Table 27) that the variance ratios are greater for the rating scales.

1.3. The skill-specific rating scale was internally consistent across both cases (α=.84 and .86 for Depression and Delirium respectively) with ITC=.45 and .37 as evidence of discrimination. These values were higher than for the two checklists (see Table 27), except for the ITC for Delirium, which was almost the same as for CL-2 (ITC=.37 versus ITC=.38). Further, the test taker variance components (also in Table 27) were larger for the rating scales than for the checklists. Figure 5 shows (as does Table 27) that the variance ratios are larger for the rating scales but the difference is less than was seen with the case-specific rating scales.

1.4. Cut Scores, Φ-coefficients and Pass Rates (as shown in Table 12)

1.4.1. Cut scores for the study instruments were lower than for the checklists with one exception, CRS-2 (57.7%) versus CL-2 (51.5%).

1.4.2. Φ-coefficients were always higher for the study instruments.

1.4.3. Borderline groups were smaller and pass rates higher for Depression than for Delirium, regardless of the scoring instrument.

2. Logistic Regression Analyses

2.1. The three-predictor model significantly fit each data set but association of the predictors to the pass-fail decisions was not strong (i.e., maximum explained variance was 16.4%, see R_L^2 values in Table 27).

2.2. R_L^2 values were always higher for the study instruments than for the checklists.

2.3. The regression model more accurately predicted the pass-fail decisions for the case-specific rating scales than for the checklists (Tables 13-18). The convergent correlations for Study One were lower than the discriminant correlations (Table

19), raising a question about the possibility of overlap between the dimension being assessed by the case-specific rating scales and that being measured by the PIRS.

2.4. The regression model did not predict the pass-fail decisions for the skill-specific rating scales any more accurately than it did for the checklists (Tables 20-25). However, like Study One, the convergent correlations for Study Two were lower than the discriminant correlations (Table 26) and showed that there was a strong correlation between the skill-specific rating scale and the PIRS.

3. Comparing Across Studies
   3.1. Delirium is the more difficult case based on the item analysis (e.g., lower mean scores and lower pass rates across all instruments) as summarized in Table 27.
   3.2. The α- and Φ-coefficients are always higher for the study instruments (Table 27).
   3.3. The regression model fit the CRS-1 and CRS-2 data best as seen by comparing results from Tables 13 to 18 to Tables 20 to 25 or by comparing R_L^2 values and the odds ratio for Skill as reported in Table 27.
4. Alternate Expertise Models
   4.1. Although different approaches to representing Experience and minimizing irrelevant variance in Knowledge were explored, none of the results (reported in Appendix K) were an improvement on the original model.
5. Assessor Effect
   5.1. Assessor effect was controlled for in the regression analyses to minimize the possibility of rater differences masking instrument differences.
   5.2. Assessor Feedback
      5.2.1. Assessors did not show a clear preference for any one type of instrument.
      5.2.2. Rating scales take more time to complete and within the OSCE used for this research completing them on time was sometimes challenging.
      5.2.3. Assessor comments and ratings indicate that the assessors found the skill-specific rating scale more troubling than any other instrument.

The remaining challenge is to consider the implications of the results within the context of the research questions and the possibilities for future work.

Discussion

This research asks whether case-specific rating scale scores and pass-fail decisions are better than checklist scores and pass-fail decisions at representing a professional competency and then asks the same question regarding skill-specific rating scale scores. Data for the two studies were collected from test takers during an OSCE that is a prerequisite for medical licensure. Discrimination and reliability of both the scores and the pass-fail decisions from each study were the criteria for the comparisons. Clinical reasoning models and questions from previous studies contributed substantially to the framework for the two studies.

The purpose of the research was to assess scoring alternatives to the checklists commonly used by OSCE raters. Checklists measure the quantity of relevant information a test taker elicits from the patient. This approach to scoring has been criticized for being too detailed and therefore being an inappropriate measure of clinical expertise because the scores do not represent qualitative indicators such as efficiency. Raters for the OSCE in this research used two scoring instruments. One was a case-specific checklist. The second was the PIRS, a rating scale that measures behavioural aspects of the test taker - patient interaction with items for behaviours like "initiation of interview", "questioning skills", and "establishing rapport".

One alternative to checklists, as explored by other researchers, is a generic rating scale. A generic rating scale is of interest because the format focuses raters on qualitative judgments about dimensions of clinical competence rather than relying on the detailed procedural criteria of case-specific checklists. A generic instrument has the further

advantage of being useful across multiple patient cases, thereby reducing development costs. However, studies to date have not been able to clearly identify whether the intended dimensions of clinical competence are represented by scores from a generic rating scale.

The rating scales developed for this research were intended to address the shortcomings of checklists while retaining the advantage of more specific criteria than is possible with a generic instrument. In Study One, case-specific rating scales with items drawing on key features of each case were used. In Study Two, skill-specific rating scales were used with items based on critical aspects of history-taking.

Details of the answers to the research questions are given in the Results chapter and are summarized in Table 27. The scores from the case-specific rating scales in Study One are more internally consistent, the variance ratios are larger, the cut scores are more dependable and the regression model explains more of the variance in the pass-fail decisions than it does for the checklists (although only by a small percentage). In short, the results provide support for the argument that the case-specific rating scale scores and pass-fail decisions discriminate more accurately among test taker performances than do checklist scores and pass-fail decisions within a construct of clinical expertise. The results also show that case-specific rating scale scores and pass-fail decisions are more reliable and dependable than those from the checklists. The results from Study Two are essentially the same. The data provide support for the argument that the skill-specific rating scale scores and pass-fail decisions discriminate more accurately than do the checklist outcomes within a clinical expertise construct. Further, the skill-specific rating scale outcomes are more internally consistent and dependable than the scores and pass-

fail decisions from the checklists. Results of previous studies are congruent with the results from Studies One and Two, although earlier work only considered scores, not pass-fail decisions. However, when case differences, assessor feedback, and correlations to PIRS scores are considered, more substantive answers emerge.

The combined results of the two studies present a curious array of data. Although there are indicators that rating scale scores are more discriminating and consistent than checklist scores, there are also indicators that the dependability of the rating scale decisions may have been achieved at the expense of construct validity. Correlations between the rating scale scores from both studies and the PIRS scores are higher than the correlations between the rating scale scores and the checklist scores, suggesting that assessors using the rating scales may focus too narrowly on those aspects of performance already being measured by the PIRS rather than on the key features dimension they are expected to judge. Concurrently, the ITC values for the rating scale scores are strong, indicating the rating scale scores correlate well with the total OSCE score and therefore are evidence that the rating scale scores do represent the construct of clinical expertise more broadly. To address this apparent contradiction the difference between the two cases (Depression and Delirium), the assessors' perceptions and cross-study comparisons are examined. Then comparisons between instruments and conclusions regarding the research questions are explored.

External Factors

Case Effect

Two external factors influencing scores in these studies are considered; specifically, case effect and assessor effect. For this research, a case is the combination of

the test taker's task and the patient's problem. The two cases selected for this research, Depression and Delirium, were as closely matched as possible, but the match was not perfect. The Delirium case is a complex task and is arguably a more authentic case by Wiggins' (1989) standards because it more closely represents the challenge a physician would face in practice, specifically assessing and managing the patient.

The Depression case is a single task of eliciting a relevant history from the patient. With this case, the task is built onto a realistic patient simulation but represents a more limited "biopsy" of the physician's total task of assessing and managing the patient. This makes the gap between the case scores and actual practice wider for Depression than for Delirium and makes the scores less comparable.

What is the effect of this case difference? The more authentic case, Delirium, is the more difficult case, as evidenced by lower mean scores (see Table 27), more test takers rated as borderline and lower pass rates (see Table 12), regardless of the scoring instrument. Assessor comments confirm that Delirium is a more difficult case, at least in part because of the time limit; for example, "...very complex station for only 9 minutes" and "The concept was good ... but there were too many things for the candidates to accomplish in the allotted 10 minutes." Although the difference in complexity and difficulty between Depression and Delirium complicates the comparison of scoring instruments, it was not sufficient to conceal instrument differences.

Assessor Effect

The second external factor to be considered is assessor effect. In the regression analyses, rater effect was minimized by selecting only those records where there was consensus on the PIRS. Another perspective on assessor effect comes from assessors'

feedback about the rating scales. Assessor perceptions have not been reported in previous studies about OSCE scoring instruments. In this research, assessor perceptions became a critical component to understanding how the rating scales were being interpreted and how modifying the rater task within the structure of a particular OSCE impacts the feasibility of implementing a new scoring instrument. For instance, in this OSCE assessors had ten minutes while observing a test taker and two minutes afterwards to complete the scoring task. The assessors who first tested the rating scales noted that the time allowed was problematic, as did assessors from the OSCE (Appendixes L and M).

Other high stakes assessments implicitly or explicitly acknowledge the scoring/time conflict with extensive rater training (Centrex, 2006a), by not regulating the time allowed for scoring (Wolfe & Gitomer, 2001) or ensuring that the design of the mark sheet makes the scoring task as clear and quick as possible. For instance, with some forms of pilot testing, the rater must run the flight simulator while observing and scoring the flight crew's performance (Brannick, Prince, & Salas, 2002) so easy-to-score instruments are especially important. Increasing the time for scoring seems the obvious solution for an OSCE but adding even small amounts of time almost inevitably decreases the number of cases or performance samples that can be included, which in turn limits the validity of the OSCE. So, like the pilot testing researchers, OSCE test developers need to be sensitive to the interaction between the time factor and the scoring task for any given instrument.

What is particularly interesting about the assessor feedback is the apparent contradiction between their expressed view that the rating scales are subjective, implying the scores are not reliable, and the data indicating that the rating scale scores are

internally consistent and correlate well with total examination scores. Although the statistics point to the rating scales as being reliable, 9 of the 58 assessors (from a total of 190) who gave comments specifically describe the rating scales as "much more subjective", implying the scales are not reliable. Nine more assessors refer to the issue of subjectivity obliquely as they describe their struggle to assign a rating when their judgments do not fit the rating scales. For example, an assessor who scored the CRS-1 reports that "... the anchors didn't really correspond always with the rating (examinee could meet the anchor but wasn't really excellent)" and another assessor, who scored Depression with the skill-specific rating scale, wrote "some people were excellent and may have missed one or two rating areas completely".

These two assessors represent both ends of a continuum. They both speak as if they scored as best they could with the given rating scale items and anchors. The first assessor states this led to inflated scores and expresses a concern about the risk of generic rating scales, namely that case-specific criteria are minimized and those test takers with poor clinical reasoning may still score moderately well. The second assessor, from scoring the more generic rating scale, expresses the concern that excellent performance is not being recognized. Implicit to his statement is the belief that the rating scale items do not incorporate the personal criteria he believes define excellent performance. Taken in conjunction with statements from two other assessors, I think the second assessor's personal criteria had a case-specific component not represented by the more generic skill-specific rating scale items. Making this hypothesis is in keeping with clinical reasoning theory which contends that the assessor, an experienced physician, most likely thinks in terms of case-specific criteria (Norman, 2005b)

The comments from the two other assessors explicitly describe the scoring challenge in case-specific terms and indirectly raise the issue of subjectivity, based on their experience with CRS-1. The first of these two assessors writes that it is "...more difficult to assess candidates with rating scales (in terms of knowledge base)" while the second one recommends that the "...checklist would be more useful to acknowledge physician's acquisition of history related to symptoms – rating scale didn't allow for accurate assessment of this". These assessors are complaining about lack of clarity related to points they consider critical to the history-taking task based on their definition of the key features of the case. Although the case-specific rating scale anchors do include aspects of the key features, it is not sufficient to satisfy these assessors' desire for case-specific criteria. In fact, for Study One only five (11%) of the CRS-1 assessors thought the rating scale comprehensively captured performance and just three (6%) of the CRS-2 assessors thought the same. For Study Two, only one assessor (1%) across both cases thought the skill-specific rating scale comprehensively captured performance. (See Appendixes L and M for more details.)

The assessor feedback does not indicate a clear preference for either rating scales or checklists. However, scoring a rating scale is perceived as a more demanding task than scoring a checklist. This was reported by pre-test assessors and their perception is echoed by comments from 8 of the 58 assessors who provide comments. Scoring a complex performance with a six-item rating scale is demanding in part because it forces assessors to fit their observations into a limited number of abstract categories (Marbry, 1999), and in this OSCE, to do so in a very short amount of time. The limited number of rating scale items limits the scope of variability. The concern is that the limited scoring options

promote inter-rater agreement and more consistent scores by narrowing the scope of what is assessed. The negative consequence of this effect is a serious potential loss of content relevance (Marbry, 1999) that could limit the validity of any inferences (Nichols & Sugrue, 1999).

The difficulty some assessors experience may be cognitive dissonance triggered by observing test taker performance that cannot be reconciled easily with the rating scale criteria. Consciously or unconsciously invoking internalized personal criteria to resolve the dissonance would simplify the scoring task and simultaneously provoke criticism about the subjectivity of the instrument that forced the response. As discussed by others (Hunter, Jones, & Randhawa, 1996; Reilly, Henry, & Smither, 1990), invoking personal criteria to reinterpret or simplify the judging process would still generate reliable scores if the assessors are falling back on criteria drawn from their shared experience as professionals.

With an OSCE, the corollary to findings in personnel studies like Reilly's (1990) would be the hypothesis that test taker behaviours that assessors perceive as desirable in individuals about to become their colleagues are being used as ancillary criteria for assigning rating scale scores. There are many desirable attributes not directly related to assessing and managing a patient that are not included in the rating scales; for example, confidence, language skill, and ease with the patient. These are desirable characteristics in a physician that likely increase with experience and developing expertise (Hodges & McIlroy, 2003); they are also attributes that are subsumed by the behaviours assessed with the PIRS. Although these attributes are not direct indicators of clinical reasoning, they likely correlate with developing clinical expertise because they are skills that also

grow with experience. If assessors are resolving a dissonance created by the need to simplify complex observations down to six ratings by invoking such criteria, then the correlation between the study instrument scores and the PIRS scores would be explained, as would the ITC. The use of subjective criteria that correlate with developing expertise and that are drawn from a common professional culture would produce internally consistent scores. This explanation would also reconcile the disparity between the assessors' perception of rating scale subjectivity and statistics indicating internally consistent rating scale scores.

Consideration of case effect and the assessor feedback provide critical context to the comparison across instruments which follows and details are integrated into the ensuing discussion.

Instrument Effect

Rating Scales versus Checklists

The rating scale scores differ from the checklist scores in several fundamental ways and details are provided in Table 27. First, the checklist scores are multidimensional (based on PCA) with relatively low internal consistency (range for α is .47 to .64), meaning the dimensions are not strongly correlated. The rating scale scores are unidimensional (based on PCA), which results in high internal consistency (range for α is .79 to .87) and means the rating scale scores rank test takers more reliably than do the checklist scores. High internal consistency also means that rating scale items are inter-correlated and homogeneous; they are measuring the same dimension, as intended. Second, there is less overall variance in the checklist distributions (as indicated by the variance ratios shown in Figures 4 and 5 and reported in Table 27), which means

checklist scores do not discriminate as accurately as rating scale scores. Third, the gap between the mean score and cut score is always greater for the rating scales, which contributes to the dependability of their cut scores and their consistently higher pass rates (i.e., Δ=14%, 8%, 25% and 17% for CRS-1, CRS-2, SRS-1 and SRS-2 respectively; Δ=9%, 5%, 8%, and 6% for CL-1 CL-2 (Study One), CL-1 and CL-2 (Study Two) respectively, as derived from values reported in Table 27). Fourth, the regression model fits the rating scale pass-fail decisions better than it fits the checklist pass-fail decisions, even though it explains only a small percentage of the variance for any instrument. However, the weakness of the model and the difference in the dependability of the cut scores between the checklists and the rating scales means that the multiple correlations might have been underestimated.

As described in the Methodology chapter, there is indirect evidence of content validity from experts for all the instruments and as shown in the Results chapter, there are modest but significant correlations between checklist and rating scale scores for each case (as shown in Table 27). The evidence is that all six instruments (i.e., two checklists, three rating scales and the PIRS) assess the same construct of clinical expertise. However, the difference in dimensionality between the checklist scores and the rating scale scores indicates that either these two methods are measuring different dimensions of the same construct (e.g., key features versus patient interaction behaviours) or they are measuring the same dimension differently. The intent of the research was to compare two approaches to scoring (i.e., checklists versus rating scales) that measured the same dimension differently in the interests of developing a superior instrument to either a checklist or a generic rating scale.

A difficulty with the checklists is having one score represent multidimensional items, which masks what is being measured and lowers the internal consistency of the score. Also, checklists are about the quantity of relevant information gathered. They do not take into account the judgment exercised in the gathering of the information. By contrast, the rating scales were meant to represent the judgment exercised relative to the importance of the information gathered by a test taker. The relatively high correlation between the rating scale scores and the PIRS scores (range of r is .68 to .82) as compared to the correlations for the checklist scores and the PIRS scores (range of r is .46 to .54) is interesting because the rating scales and the PIRS were supposed to measure two different, albeit related dimensions of test takers' clinical expertise: Efficiently eliciting key features of the patient problem and effectively interacting with the patient. While the rating scale scores are internally consistent, strong correlations with the PIRS scores makes it unclear as to how well the rating scales used in this research are measuring the key features dimension of history taking.

A certain degree of correlation with the PIRS scores is expected for the case-specific and skill-specific rating scales, as well as for the checklists. All these instruments and the PIRS are intended to measure dimensions of the same clinical expertise construct and as such, there is an assumed positive relationship among them. The ITC values for the PIRS are high and as noted in the Results chapter, approximately 15% of the total OSCE score is based on PIRS scores from 5 of the other 12 cases, so this is not surprising. The combination of strong correlations between the case-specific rating scale scores and the PIRS scores (r=.8 and r=.7 for Depression and Delirium respectively) and strong ITC values for all these instruments (ITC for CRS-1=.48 and CRS-2=.39; ITC

range for PIRS: .49 to .51) suggests that the case-specific rating scale scores may represent a dimension very closely aligned with the PIRS, if not the same dimension. The same result is seen with the skill-specific rating scale scores which correlate to the PIRS scores by .7 for both cases with high ITC values for all (ITC for SRS-1=.45 and SRS-2=.37; ITC range for PIRS: .38 to .49), leading to the same concern that the skill-specific rating scale scores may represent the same dimension as the PIRS.

The PCA (Appendix G) reflect the same rating scale-PIRS relationship. In Study One the PCA extracted one component from the combined CRS-1 and PIRS items and items from both instruments contributed to that component; in other words, one dimension was assessed by CRS-1 and PIRS. For CRS-2 and both the skill-specific rating scale data sets, the same analysis led to two components being extracted. However, all the items from each rating scale and the PIRS contributed to the first component and some items from CRS-2 or the skill-specific rating scale contributed to a second component. Although two components were identified, the PCA did not appear to distinguish a component represented by PIRS as separate from one represented by any of the study instruments.

PIRS scores are expected to correlate strongly with overall OSCE performance but they represent only one dimension of the clinical competence construct. The rating scales from Study One and Study Two were not intended to be extensions of the PIRS; they were supposed to represent the key features dimension. According to the study data, this expectation is not met. Given that the case content has been validated by experts and by performance in previous OSCE administrations, a likely explanation for the

discrepancy between expected and actual results may be the impact of the scoring methodology on the assessors.

The argument that checklists reward thoroughness, thereby disadvantaging more expert and therefore more efficient test takers was formally introduced in the 1999 study by Hodges and his colleagues. However, they compared generic rating scale scores to scores from checklists intended to assess medical students. Subsequently, Hodges and McIlroy (2003) cautioned that scoring instruments that are valid for one level of competence such as novice medical students may not be valid for another level, such as expert doctors with several years of clinical experience. In keeping with this caution, the checklists for the OSCE used in this study were designed to assess history taking by post-graduate trainees seeking medical licensure. The checklist items were limited to information deemed critical to the task and patient problem with the intention that efficient test takers with the appropriate clinical focus would be better rewarded than would novices with a dogged step-by-step approach or test takers who simply ask a barrage of questions hoping to get enough right. Even when compared to the checklist scores designed to reward efficiency, rating scale scores proved more reliable and discriminating, providing a robust replication of earlier findings. The conclusion that checklists disadvantage test takers with more clinical expertise is not so clearly supported by this research because there are no clearly identified experts in the cohort. None of the internationally trained clinicians with experience in independent practice were from the two disciplines most relevant to the two patient cases used in the studies; namely, psychiatry or family medicine. The Canadian-trained test takers were all trainees without experience in independent practice.

What this research does suggest is that checklists may disadvantage less expert but nonetheless competent test takers because the time limit emphasizes the need for efficiency. For example, a thorough test taker who pursues the patient history methodically but in unnecessary detail due to lack of experience and expertise will run out of time before progressing through all the critical aspects of the patient's problem, even if she is competently working her way through it. In this case, the lower score may not indicate incompetence, just less expertise. Discriminating between the thoroughness of this individual and someone who simply amasses a quantity of information is a matter of judgment that takes into account factors like the number of irrelevant questions asked and the order in which information is gathered (Hodges, McNaughton, Regehr, Tiberius, & Hanson, 2002); that is, factors that are not accounted for by a checklist. The competent but less expert test taker might have a low checklist score due to lack of efficiency but do quite well on the PIRS behaviour items for being organized, using open-ended questions well and developing a rapport with the patient.

The problem is that examiners using a checklist may feel compelled to score a checklist item whenever a relevant question is posed. They may do so regardless of whether the test taker takes a rational and competent approach or takes a "shotgun" approach to eliciting information from the patient, and despite the expectation in the OSCE used for this research that only items completed *satisfactorily* should be scored. Therefore a test taker who simply asks a lot of questions may achieve the same score as a more competent novice, but more as matter of chance than method. In effect, checklists designed to assess more experienced test takers minimize discrimination between competent novices and rapid-fire but poor performers by rewarding quantity, not

thoroughness. The result is a restricted range of scores at both tails of the checklist distributions, which are truncated relative to the rating scale distributions.

The disparity between the checklist and the rating scale score distributions is most evident when lower scores are compared. Since the lowest checklist scores are not that low, we know that all the test takers elicited at least a modicum of relevant information, regardless of their understanding of the patient problem. With the rating scales, however, in each case and both studies, some test takers received very low scores. This clearly suggests that the rating scales produce scores that measure something other than the simple quantity of information gathered. As well, many test takers achieve very high scores with the rating scale, suggesting that excellent test takers earn a perfect score even though they do not elicit every bit of relevant information (using the checklist items to define the relevant information). As a result, there is more variance in the rating scale scores compared to the checklist scores and the test taker variance components are larger. This finding indicates greater discrimination across test takers and is evidence for a construct validity argument in favour of rating scale scores.

Whether the more reliable scores from the rating scales are achieved by inappropriately narrowing the assessors' perception of the construct remains an open question. The strong correlation between the rating scale scores and the PIRS scores, in conjunction with the PCA, ITC values and content validation, indicates that case-specific and skill-specific rating scales may be measuring the same dimension as the PIRS, or one very closely related to it. One possibility is that the rating scale format may be a more significant factor in determining what is assessed than the dimensions described by the anchors for the items within each rating scale, perhaps because the case-specific rating

scale and skill-specific rating scale anchors are not sufficiently defined or perhaps because the method does not match the domain.

Nunnally's (1994) discussion of methodological heterogeneity is a reminder that methodologies also have domains that should be specified. The use of rating scales for scoring behavioural observations in psychology is well established. The use of rating scales for scoring the content or key features dimension of clinical competence for educational assessment is not so clearly established. In this instance, the methodology may be determining what is scored more than the item anchors themselves. The correlations with PIRS are evidence of convergent validity, as would be expected of a monotrait-monomethod correlation, rather than evidence of discriminant validity based on a heterotrait-heteromethod correlation, which was intended. Although the different instruments (PIRS versus case-specific rating scales or the skill-specific rating scale) were expected to measure different domains, the differences between the rating scales and the PIRS may not be sufficient to achieve this goal.

If rating scales narrow the definition of the clinical expertise construct, then the risk of losing content relevance is greater for Delirium because of its greater authenticity and capacity to support inferences to clinical practice. Delirium is the case where the most information is lost if assessors view test taker performance through a narrow lens. As the more complex case, Delirium is also the case most likely to generate cognitive dissonance among the assessors as they score the test takers with "narrow" rating scales.

If rating scales better represent clinical expertise and the Delirium case is a more authentic task, then the rating scale scores for Delirium should be more discriminating than for Depression. Instead, the α-coefficient for CRS-2 is significantly lower than the

coefficient for CRS-1 (based on non-overlapping confidence intervals shown in Tables 7 and 8) and the test taker variance component is also smaller for CRS-2 than for CRS-1 (Δ=12.4%, as taken from values in Table 9). CRS-2 scores are *less discriminating* than CRS-1 scores, contrary to what was just hypothesized. With the skill-specific rating scale, the difference between cases is less. The α-coefficient is virtually unchanged between SRS-2 and SRS-1 (as shown in Table 10) and the test taker variance component for SRS-2 is only larger by 4.1% (as taken from values in Table 11). The skill-specific rating scale *discriminates about equally* between the two cases.

Conversely, checklist performance is better with the more complex Delirium case. The α-coefficients for the CL-2 are significantly higher than the coefficients for CL-1 in Study One and the difference found in Study Two is also significant. The test taker variance component for CL-2 is higher than for CL-1 by 0.9% and 1.2% for Studies One and Two respectively (derived values in Tables 9 and 11). Although the increases in the test taker variance components for the checklists appear small, they are proportionally on the same scale as the change reported for the case-specific rating scales. Based on this criterion, CL-2 is *more discriminating* than CL-1.

Is there a pattern related to content relevance in these differences? Insofar as there can be a pattern across two cases, the case-specific rating scales are sensitive to case differences but represent the more authentic case less accurately, counter to the rationale for using a rating scale. However, a rating scale may be a better method of scoring for a case like Depression where the history-taking task is arguably more dependent on the rapport established with the patient. The rating scale methodology may not be as suitable for scoring the Delirium case where the task requires more time management, directive

interviewing and diagnostic judgment. There is little difference in discrimination across case with the skill-specific rating scale, which runs counter to assumptions about the value of using generic rating scales to assess clinical competence. The lack of an expected response to the case differences added to the pattern of high correlations and ITC values (Table 27) further suggests that the case-specific and skill-specific rating scales may be extensions of the PIRS. If true, this is evidence that a loss of content relevance has occurred because the case-specific and skill-specific rating scales are not discriminating along the expected key features dimension currently represented by the checklists.

Looking more closely at the variance ratios (Table 27) provides an interesting view of the case-based comparisons. The ratios for CRS-2 and for SRS-2 were larger than for their Depression counterparts. However, the test taker variance component for CRS-2 was smaller than for CRS-1 (also in Table 27), indicating that the increase in the variance for CRS-2 was error related. Of interest, the increase was related to relative error as it occurred in the test taker-item (*tti*) variance component, not in the item (*i*) variance component (see Table 9). For the SRS-2 the reverse was true as the test taker variance component was higher than for SRS-1, but not by much, and only because of a shift in variance from the test taker-item variance component (Table 11). The test taker-item variance component may not be related to relative error variance (Nichols & Sugrue, 1999) if the clinical expertise construct explains and predicts interactions between test takers and items when items differ in their cognitive demands and test takers differ in their approach to those demands. Factors influencing test takers' approach would include their educational background, clinical specialty and professional experience. If test taker-

item interaction is explained by the construct, it is not contributing to relative error variance. Unfortunately, the clinical expertise construct is not well enough developed within this research to determine whether the test taker-item variance component is or is not related to relative error variance. All the same, the possibility that this variance component does not represent relative error is thought provoking because the difference in variance components between the checklists and rating scales is considerably less if error is restricted to the item component. Treating the test taker-item variance component as construct-relevant variance would provide a stronger validity argument for checklist scores than is supported by a more traditional interpretation of the variance components.

The dependability of the rating scale cut scores is a positive characteristic. However, the dependability is in part due to the distance between the cut scores and the mean scores, a distance that is consistently greater for the rating scales across case and study. Both cases are also easier for test takers when scored with a rating scale. The rating scales from both studies have higher mean scores and higher pass rates than occurs with the checklists. Are the cases less difficult because rating scale scores represent a narrower definition of the construct than anticipated? Is a more dependable cut score worth a loss of score validity? The second question is harder to answer and goes beyond the purview of this research. The answer involves a judgmental trade-off that depends on the consequences of decision error and the severity of the loss of validity relative to the objectives of the OSCE as a whole. What this research does compare is the relative validity of the pass-fail decisions across instruments based on the regression model for clinical expertise, which provides a start to answering the first question.

The regression model, essentially represented by the Skill variable, explains more of the variance (as indicated by the R_L^2 ratios reported in Table 27) in the rating scale pass-fail decisions than it explains for the checklist pass-fail decisions. In general, this result indicates the rating scale decisions are more valid because the model fits them better than it fits the checklist decisions. However, as noted earlier, taking into account the weakness of the model and the difference in the dependability of the cut scores between the checklists and the rating scales suggests that the model fit may owe more to the reliability factors than to expertise predictors. The most reliable predictor more accurately predicts the more dependable decisions. Whether the model fit is interpreted as an indicator of validity or reliability, evidence of significant model fit supports the proposition that the rating scales are measuring the same construct as the other OSCE cases, which are represented by the Skill score in the model. However the regression analyses do not clarify which dimensions of the clinical competence construct are being measured by the checklists or the study rating scales and so these broad comparisons leave the first question unanswered. Looking more closely at the two types of rating scale is the next step toward an answer.

Case-specific Rating Scales versus Skill-specific Rating Scale

So far this discussion has focused on broad comparisons between rating scales and checklists. However, comparing the case-specific and the skill-specific rating scales is also instructive and moves the discussion closer to answering the research questions. In some respects, as implied by the foregoing discussion, there is little difference between the two types of rating scales. For example, the case-specific rating scale cut scores are about as dependable (Φ=.86 and Φ=.79 for Depression and Delirium respectively) as the skill-specific rating scale cut scores (Φ=.84 and Φ=.89 for Depression and Delirium

respectively). Furthermore, the case-specific rating scale scores correlate with the checklist scores (r=.30 and r=.50) to about the same degree as the skill-specific rating scale scores do (r=.36 and r=.41).

Where a difference between instruments does occur is in the regression analyses as the model explains slightly more of the variance in the case-specific rating scale pass-fail decisions than it does for the skill-specific rating scale pass-fail decisions. For example, for Depression, $R_L^2{}_{CRS\text{-}I}$= 16.4% versus $R_L^2{}_{SRS\text{-}I}$ = 12.7%. Although this difference is not large, it occurs with both cases showing that the expertise model, largely represented by Skill, fits the case-specific rating scale decisions better. This finding may be interpreted as stronger support for the validity argument for the case-specific rating scales; an interpretation that is not attenuated by any significant difference in cut score dependability.

Assessor comments are another sign of difference between the two types of rating scales. The skill-specific rating scale, despite anchors intended to capture elements of clinical reasoning, is a generic, skill-based instrument. The assessors give the skill-specific rating scale lower ratings for comprehensive coverage, make more comments about its subjectivity and about the need for more specific anchors than they do for the case-specific rating scale. Perhaps the strongest criticism of the skill-specific rating scale comes from the assessor who reported "The difference between [items] 1 & 2 [and] to some extent 2 & 3 were not clear. I did look at the other marking sheet to try to sort this out and it still wasn't clear. … I ended up deciding 1 = hx of hallucinations 2 = mini mental/call 3 = ddx." This assessor is so uncomfortable with the skill-specific rating scale that he tries to recreate the checklist within the rating scale format. Another skill-specific

rating scale assessor, who also scored Delirium, comments that "the rating scale should be less generic and more customized to the content of the station", providing anecdotal support for the case-specific rating scales. The assessor ratings indicate a lack of comprehensive coverage for both types of rating scales (Appendices L and M) and their comments reveal uneasiness about what they are assessing with the skill-specific rating scale.

Despite assessors' uneasiness, the skill-specific rating scale scores are still internally consistent and the skill-specific rating scale pass-fail decisions are very dependable. The internal consistency of the skill-specific rating scale scores reinforces the possibility that the assessors may have overcome the instrument's shortcomings by falling back on some form of implicit, internal criteria common to the profession. However, the differences between the cut scores and the mean scores for the skill-specific rating scale are almost twice the differences seen with the case-specific rating scale; specifically, for CRS-1 the Δ=14.6% and for CRS-2, Δ= 8.6%, whereas for SRS-1, Δ=25.2% and for SRS-2, Δ=16.5%. The cut score is the mean of test takers rated as borderline. If the relatively low cut score is a consequence of assessors' uneasiness about scoring, then their reluctance to rate test takers as even borderline satisfactory is a negative consequence because it effectively lowers the pass standard. Furthermore, the low cut score suggests that cases may be less difficult because of how the rating scale represents the construct to the assessors, although perhaps more because of ambiguous items than a narrow perspective.

Overall, the results in Table 27 show that the rating scales do provide for more discriminating and reliable pass-fail decisions for a dimension of the clinical expertise

construct, but it is a dimension closely correlated with the behaviours assessed by PIRS. Furthermore, these data demonstrate that the case-specific rating scales provide a better balance between reliability and validity than does the skill-specific rating scale. Whether either form of rating scale is measuring the intended dimension of clinical reasoning is unclear. If the rating scales under-represent the key features dimension then there may be a serious loss of content relevance.

Conclusion

This was a large-scale, comprehensive comparison of scoring instruments within the demanding context of a high stakes performance assessment of professional competency. In this context, the need for consistent and discriminant pass-fail decisions, not just scores, is critical. The negative consequences of poor decisions impact everyone, from the test taker who should have passed, to the public who are served by test takers who should have failed, to the test developer who is liable for both types of error. Because of the potential consequences, a more intensive look at the type of rating scales that might be suitable for scoring high stakes performance assessment of professional competency was needed. Hopefully, the detailed examination of the two types of rating scales considered in this research will provide direction to test developers and researchers.

The Study One results add to the discussion with the conclusion that *case-specific* rating scales are more reliable and discriminating than checklists for scoring a high stakes performance assessment, with one critical caveat. The results do not clarify to what extent the rating scale methodology is capturing the intended key feature dimension of clinical competence. For some OSCE cases, the task is strongly linked to the PIRS

dimension. For example, with Depression successful history-taking may rely as much on the nature of the interaction and the rapport with the patient as it does on clinical reasoning skill. The data support the use of a case-specific rating scale for these cases. However, these data do not provide the same degree of support for a case-specific rating scale when multiple dimensions of clinical competence, like assessment skills and the ability to make treatment decisions, are being assessed. In this respect, Study One adds force to the earlier conclusions about the need to clarify what is being assessed by any one rating scale before this methodology is used extensively for high stakes assessments.

Study Two leads to the conclusion that *skill-specific* rating scales share the limitation given for generic rating scales by Reznick and his colleagues (1998), namely that while the scores are reliable and valid (based on indicators of discrimination and the item-total correlations) it is not clear that the intended dimension of clinical competence is well represented. Although the data from Study Two show a correlation between the skill-specific rating scale scores and the PIRS scores, the case-specific rating scale scores are psychometrically stronger. However, the skill-specific rating scale data are more than a repeat of previous work. The important contribution of Study Two is the additional evidence it provides about the correlation between rating scale scores and PIRS scores found in Study One. This finding is a step forward in clarifying what dimension is being assessed by rating scales and is a cautionary message to OSCE developers.

Taken as a whole, this research illustrates how comparing checklists to rating scales can create a false dichotomy that oversimplifies the validity issues by focusing on the methodology almost to the exclusion of case factors. Based on the data in these two studies, case-specific rating scales may have psychometric (given the comparison to the

more generic skill-specific rating scale) and theoretical advantages over generic rating scales and deserve further development. The data also suggest that case-specific rating scales may not be the best method for all types of OSCE cases.

Dichotomous comparisons based on methodology are also prone to neglecting the interaction between assessors and the scoring format because they ignore the potential interaction between case complexity and the scoring process. Checklists are easier to score, a factor which should not be ignored but one that may not have arisen with studies (e.g., Hodges & McIlroy, 2003; Regehr et al., 1998) where assessors scored a checklist first. Scoring a checklist first mimics Reilly's (1990) use of checklists as an aid for developing rating scales and using Reilly's logic, scoring a checklist first minimizes the dissonance assessors may experience when working with more abstract scoring criteria and therefore lessens the time needed to score. This advantage is lost when the rating scale is scored without benefit of the checklist information.

The wish to replace case-specific checklists with a generic rating scale is strongly motivated by the gain in utility that can be realized. If OSCE developers step back from the level of utility offered by a generic scoring instrument, then other scoring possibilities become apparent. The empirical evidence from these two studies should move OSCE developers toward wider implementation of rating scales. However, the results of these two studies also recommend that a case-based approach to developing new instruments be continued.

Limitations

There are some important limitations to the research. Generalizability of the findings is limited by the nature of the OSCE and of the two cases. This was a physician-

scored OSCE and the results may not generalize to a patient-scored OSCE where the rater is not a clinical expert. Despite data being collected from a multidisciplinary OSCE, a move much advocated by other researchers, the two cases that best suited this study are both drawn from psychiatry. Since the results suggest that the interaction with case is an important factor, it is somewhat unclear to what degree the results generalize beyond the psychiatric domain. That being said, both cases are representative of common and important patient problems presented in contexts (e.g., first presentation of depression at a clinic, delirium as post-operative complication in a hospital ward) where the discipline of the test takers should not have been a deciding factor in providing initial care to the patient. The critical difference between these two cases was lodged more in the complexity of the task than in the nature of the patient problems and therefore there are solid grounds for generalizing beyond psychiatry.

The simplicity of the regression model is a third limitation. The model only explained a small proportion of variance in the pass-fail decisions, thus limiting the weight that can be placed on the conclusions regarding the discrimination of the pass-fail decisions. Improving the model would require identifying more and stronger predictor variables. Practically speaking, this is a real challenge when focusing on professional competencies. A more productive approach might be to study experts by including them with a cohort of actual test takers, thus combining the experimental strength of earlier studies with the advantages of using a high stakes OSCE for data collection. A limitation of some previous studies was the small number of participants; a limitation of this study is the nature of the cohort as there was no clear way to compare more expert to less expert test takers. Although the cohort for this research included many test takers with

multiple years of clinical practice their experience was confounded by international training, their age range, and specialties that disadvantaged them relative to these two cases. However, recruiting practicing doctors to take a licensure exam for research purposes is a daunting challenge and raises interesting ethical issues if they should do poorly.

Another improvement would be extending the study across more cases. This could be accomplished within the design of this study and would simply lead to a smaller number of test takers for each case. With more cases, the focus should be on one type of instrument, the case-specific rating scale or a derivative, with attention given to improving the items.

Even with these limitations this research contributes in several important ways. The two studies present evidence that rating scale scores are more discriminating and reliable than checklist scores, even when compared to checklists designed to assess post-graduate trainees. The caveat that rating scales may under-represent the intended construct indicates the necessity of clarifying what rating scales do assess and perhaps to reconsider what they should assess. The advantage of having independent raters who also scored a common instrument was being able to control for rater effect which in turn supported the comparison of pass-fail decisions. Examining the pass-fail decisions, especially in conjunction with the assessor feedback, was an important element to the comparisons of checklists and rating scales made in this research, especially since most OSCE are used for decision making.

While it is unclear to what degree these results would generalize to performance assessments in other professions, content specificity is not a phenomenon unique to

clinical performance assessments (Eva et al., 1998) nor are concerns regarding the negative impact of generic ratings. Therefore, conclusions related to the importance of a case-based scoring instruments and to the validity risks of generic scoring instruments may be valuable signposts to any test developer who is looking to develop the most valid and reliable scoring instruments for performance assessments of professional competencies.

Future Research

Principles and practice from different streams of work influenced this research. The key features approach (Farmer & Page, 2005) was used to describe the dimension that the checklists and rating scales should be measuring but it was not well applied to the development of the rating scales and has not been applied explicitly to checklist development. Given its compatibility with current trends in clinical reasoning theory (Norman, 2005b), an explicit application of the key features approach to creating problems for written examinations could be applied to developing OSCE cases. To date most research has attached new scoring instruments onto pre-developed cases. OSCE cases developed from key features principles might lead to more valid and reliable OSCE scores. Given the correlation between rating scales and PIRS scores, consideration could be given to a very short instrument assessing the key features dimension, perhaps with only three or four items, depending on the number of key features, with limited scoring options. For example, for Depression one such item could be "suicide risk assessment" with four scoring options: No attempt, Poor attempt, Acceptable, Excellent. Then the PIRS could be extended with one or more case-specific items if there were interaction components unique to the case. This combination of "checklist" and PIRS might be

sufficient to provide assessors with common criteria for judging performance and ensure that they record the degree to which test takers demonstrate sound clinical reasoning.

Still unexplored is whether there are systematic differences in the performance of test takers who passed when scored with one instrument but not with the other to see if the dimensions each instrument is measuring can be more clearly defined. Test taker records from this study would support such an investigation even though this study design did not. A comparison of test takers with different pass-fail results would benefit from a mixed method approach that included scores from other cases, assessor observations of the test takers as well as clinical judgments about the pattern of scores found within each instrument from each test taker.

An alternate and more radical strategy is to ignore the issue of rating scales versus checklists. Norman (2005b) concludes that the most defensible measures of clinical expertise will be based on successful patient outcomes, which start with accurate diagnoses. To a large extent clinical reasoning can be assessed with computer administered key features questions. What is not captured by a computer-based assessment is the test taker's ability to apply clinical reasoning during an interaction with a patient. In a more authentic assessment like an OSCE the test taker's ability to function professionally while reasoning through the patient's presenting problem is assessed. Norman's point is that given multiple paths to resolving any given patient complaint, trying to assess the test taker's approach to the patient is problematic and perhaps irrelevant to determining clinical reasoning ability. Under the broader umbrella of clinical competence, the quality of the interaction with the patient while the reasoning process occurs is important. Maybe scoring should only consider a test taker's top two or three

diagnoses for a case, their initial treatment or investigation orders and a PIRS rating. With this approach OSCE scores would be strongly linked to key features, as represented by diagnostic and treatment decisions, and scores would be more closely linked to the cognitive processes described by the clinical reasoning construct.

In closing, completing such a large-scale research project has been an exciting challenge. Developing the instruments and collecting the data meant working with many people who were all supportive of the project, including the test committee members and the staff at each of the 12 sites who recruited the assessors and who carefully managed a considerable quantity of extra paperwork. The complexity of working with multiple data sets and multiple analyses required patience, planning and perseverance. Sifting through data to focus on the most important results was the last and most satisfying step because the results open many doors.

Introducing change to any high stakes assessment requires extensive supporting research because of the risks associated with decision-making. This research has addressed limitations of earlier studies with its scale and the scope of the analyses and should promote more extensive and appropriate use of rating scales within high stakes performance assessments like this OSCE. The results should also provide direction to anyone developing scoring instruments for the performance assessment of professional competencies, whether in a summative assessment or in a workplace setting. Moreover, the results are a stimulus for a series of research projects that should prove interesting and informative. The completion of this research is more a step forward in the development of scoring instruments that can summarize complex performances with a minimal loss of validity and reliability than it is an ending.

References

SPSS (2004). (Version 15.1.2600 Service Pack 1 Build 2600) [Computer software]. Chicago: SPSS Inc.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing.* Washington, D.C.: American Educational Research Association.

Arnold, M. E. (1996). Influences and limitations of classical test theory reliability estimates. In Unknown (Ed.), *Proceedings of the annual meeting of the Southwest Educational Research Association* (pp. 1-30). New Orleans, LA.

Austin, Z., O'Byrne, C., Pugsley, J. P., & Quero, L. (2003). *Development and validation processes for an objective structured clinical examination (OSCE) for entry-to-practice certification in pharmacy: The Canadian experience. American Journal of Psychology* Retrieved 8-7-0006 from http://www.ajpe.org/aj6703/aj670376/aj670376.pdf .

Baker, D. P. & Dismukes, R. K. (2002). A framework for understanding crew performance assessment issues. *The International Journal of Aviation Psychology*, 12, 205-222.

Bracey, G. W. (2005). *Think about tests and testing: A short primer in "Assessment Literacy". American Youth Policy Forum, Washington, DC.* Retrieved 8-8-2005 from http://www.aypf.org/publicatons/BraceyRep.pdf

Brannick, M. T., Prince, C., & Salas, E. (2002). The reliability of instructor evaluations of crew performance: Good news and not so good news. *The International Journal of Aviation Psychology*, 12, 241-261.

Brennan, R. L. (1983). *Elements of generalizability theory.* Iowa City, IA: American College Testing Program.

Brennan, R. L. (1999). urGenova [Computer software]. Iowa City, IA: Iowa Testing Programs.

Brennan, R. L. & Johnson, E. G. (1995). Generalizability of performance assessments. *Educational Measurement: Issues and Practice*, 14, 9-27.

Bulmer, M. G. (1979). *Principles of statistics.* New York: Dover.

Centrex (2006a). *OSPRE® Assessor Course. Central Police Training and Development Authority* Retrieved 1-6-2006a from http://www.centrex.police.uk/cps/rde/xchg/SID-3E8082DF-8D2DB74B/centrex/root.xsl/1069.htm

Centrex (2006b). *OSPRE® Part II. Central Police Training and Development Authority* Retrieved 1-31-0006b from http://www.centrex.police.uk/cps/rde/xchg/SID-3E8082DF-8D2DB74B/centrex/root.xsl/501.html

Clark, L. A. & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, 7, 309-319.

Committee on Assessment and Teacher Quality (2001a). Alternative assessment case studies. In K. J. Mitchell, D. Z. Robinson, B. S. Plake, & K. T. Knowles (Eds.), *Testing teacher candidates: The role of licensure tests in improving teacher quality* (pp. 298-330). Washington, D.C.: National Academy Press.

Committee on Assessment and Teacher Quality (2001b). Improving teacher licensure testing. In K. J. Mitchell, D. Z. Robinson, B. S. Plake, & K. T. Knowles (Eds.), *Testing teacher candidates: The role of licensure tests in improving teacher quality* (pp. 147-162). Washington, D.C.: National Academy Press.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart Winston.

Cunnington, J. P. W., Neville, A. J., & Norman, G. R. (1997). The risks of thoroughness: Reliability and validity of global ratings and checklists in an OSCE. In A. J. J. A. Scherpbier, C. P. M. van der Vleuten, J. J. Rethans, & A. F. W. van der Steeg (Eds.), *Advances in Medical Education: Proceedings of the Seventh Ottawa International Conference on Medical Education* (pp. 143-145). Dordrecht: Kluwer Academic Publishers.

Dauphinée, W. D., Boulais, A. P., Smee, S. M., Rothman, A. I., Reznick, R., & Blackmore, D. E. (2000). Examination results of the Licentiate of the Medical Council of Canada: Trends, Issues and Future Considerations. In D. E. Melnick (Ed.), *Proceedings of the Eighth International Ottawa Conference - Evolving Assessment: Protecting the Human Dimension* (pp. 92-98). Philadelphia: National Board of Medical Examiners.

Delandshere, G. & Petrosky, A. R. (1994). Capturing teacher's knowledge: Performance assessment a) and post-structuralist epistemology b) from a post-structuralist point of view c) and post-structuralist perspective d) none of the above. *Educational Researcher*, 23, 11-18.

Delandshere, G. & Petrosky, A. R. (1998). Assessment of complex performances: Limitations of key measurement assumptions. *Educational Researcher*, 27, 14-24.

Dillon, G. F. & Henzel, R. R. (1999). *The relationship between amount of postgraduate training and performance on a physician licensing examination*. Montreal, Canada.

Duhachek, A. & Iacobucci, D. (2004). Alpha's standard error (ASE): An accurate and precise confidence interval estimate. *Journal of Applied Psychology*, 89, 792-808.

Eisner, E. W. (1999). The uses and limits of performance assessment. *Phi Delta Kappan*, May, 658-660.

Elstein, A. (1993). Beyond multiple-choice questions and essays: The need for a new way to assess clinical competence. *Academic Medicine*, 68, 244-249.

Elstein, A., Shulman, L., & Sprafka, S. (1978). *Medical problem solving*. Cambridge, MA: Harvard University Press.

Ericsson, K. A. & Charness, N. (1994). Expert performance. *American Psychologist*, 49, 725-747.

Eva, K. W. (2002). The aging physician: Changes in cognitive processing and their impact on medical practice. *Academic Medicine*, 77, S1-S6.

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

Farmer, E. A. & Page, G. (2005). A practical guide to assessing clinical decision-making skills using the key features approach. *Medical Education*, 39, 1188-1194.

Gitomer, D. H. (1993). Performance assessment and educational measurement. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 241-263). Hillsdale, NJ: Lawrence Erlbaum Associates.

Grand'Maison, P., Lescop, J., Rainsberry, P., & Brailovsky, C. A. (1992). Large-scale use of an objective structured clinical examination for licensing family physicians. *Canadian Medical Association Journal*, 146, 146-1740.

Haertel, E. H. (1999). Validity arguments for high-stakes testing: In search of the evidence. *Educational Measurement: Issues and Practice*, 18, 5-9.

Harden, R. M., Stevenson, M., Downie, W. W., & Wilson, G. M. (1975). Assessment of clinical competence using objective structured examination. *British Medical Journal*, 1, 447-451.

Hodges, B. & McIlroy, J. H. (2003). Analytic global OSCE ratings are sensitive to level of training. *Medical Education*, 37, 1012-1016.

Hodges, B., McNaughton, N., Regehr, G., Tiberius, R. G., & Hanson, M. (2002). The challenge of creating new OSCE measures to capture the characteristics of expertise. *Medical Education*, 36, 742-748.

Hodges, B., Regehr, G., Hanson, M., & McNaughton, N. (1998). Validation of an objective structured clinical examination in psychiatry. *Academic Medicine*, 73, 910-912.

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

Holt, R. W., Hansberger, J. T., & Boehm-Davis, D. A. (2002). Improving rater calibration in aviation: A case study. *The International Journal of Aviation Psychology*, 12, 305-330.

Hunter, D. M., Jones, R. M., & Randhawa, B. S. (1996). The use of holistic versus analytic scoring for large-scale assessment of writing. *The Canadian Journal of Program Evaluation*, 11, 61-85.

Hunter, D. M., Randhawa, B. S., & Bikkar, S. (2001). The large-scale, authentic assessment of listening and speaking as interactive communication: Issues in reliability. *Alberta Journal of Educational Research*, 47, 156.

Kane, M. T. (1997). Model-based practice analysis and test specifications. *Applied Measurement in Education*, 10, 5-18.

Kane, M. T., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice*, 18, 5-17.

LaDuca, A. (1994). Validation of professional licensure examinations: Professions theory, test design, and construct validity. *Evaluation & the Health Professions*, 172, 178-197.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher*, 29, 4-16.

Lynn, S., Bickley, P. G., & Szilagy, M. P. H. (2004). *Bate's Guide to Physical Examination and History Taking*. (8th ed.) Philadelphia: Lippincott Williams and Wilkens.

Madaus, G. F. & O'Dwyer, L. M. (1999). A short history of performance assessment: Lessons learned. *Phi Delta Kappan*, May, 688-695.

Marbry, L. (1999). Writing to the rubric: Lingering effects of traditional standardized testing on direct writing assessment. *Phi Delta Kappan* 673-679.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum Associates.

McEwen, N. (1995). Accountability in education in Canada. *Canadian Journal of Education*, 20, 1-17.

Medical Research Council of Canada, Natural Sciences and Engineering Research Council of Canada, and Social Sciences and Humanities Research Council of Canada (2003). *Tri-council Policy Statement: Ethical Conduct of Research Involving Humans.* Ottawa, Canada: Medical Research Council of Canada.

Menard, S. (2001). *Applied Logistic Regression Analysis.* (Second ed.) (Vols. 106) Thousand Oaks, CA: Sage Publications.

Mertler, C. A. (2001). *Designing scoring rubrics for your classroom. Practical Assessment, Research & Evaluation* Retrieved 6-14-2003 from http://edresearch.org/pare/getvn.asp?v=7&n=25

Messick, S. (1994). The interplay of evidence and consequence in the validation of performance assessments. *Educational Researcher*, 23, 13-23.

Miller, D. M. & Linn, R. L. (2000). Validation of performance-based assessments. *Applied Psychological Measurement*, 24, 367-378.

Miller, G. E. (1990). The assessment of clinical skills/competence/performance. *Academic Medicine, 65(supp), S63-S65.*

Miller, M. D. & Legg, S. M. (1993). Alternative Assessment in a High Stakes Environment. *Educational Measurement: Issues and Practice*, 12, 9-15.

Mislevy, R. J. (1993). Foundations of a new test theory. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 19-39). Hillsdale, NJ: Lawrence Erlbaum Associates.

Moskal, B. M. & Leydens, J. A. (2000). *Scoring rubric development: Validity and reliability. Practical Assessment, Research & Evaluation* Retrieved 12-23-2002 from http://edresearch.org/pare/getvn.asp?v=7&n=10

Nichols, P. D. & Smith, P. L. (1998). Contextualizing the interpretation of reliability data. *Educational Measurement: Issues and Practice*, 17, 24-36.

Nichols, P. D. & Sugrue, B. (1999). The lack of fidelity between cognitively complex constructs and conventional test development practice. *Educational Measurement: Issues and Practice*, 18, 18-29.

Norcini, J. J. (2002). The death of the long case? *British Medical Journal*, 324, 408-409.

Norman, G. R. (2005a). Editorial - Checklists vs. Ratings, the Illusion of Objectivity, the Demise of Skills and the Debasement of Evidence. *Advances in Health Sciences Education*, 10, 1-3.

Norman, G. R. (2005b). Research in clinical reasoning: Past history and current trends. *Medical Education*, 39, 418-427.

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

Nunnally, J. C. & Bernstein, I. R. (1994). Construction of conventional tests. In *Psychometric Theory* (Third ed., pp. 293-337). New York: McGraw-Hill.

Page, G. & Bordage, G. (1995). The Medical Council of Canada's key feature project: A more valid written examination of clinical decision-making skills. *Academic Medicine*, 70, 104-110.

Pampel, F. C. (2000). *Logistic Regression: A Primer.* (Vols. 07-132) Thousand Oaks, CA: Sage.

Parker-Taillon, D., Cornwall, J., Cohen, R., & Rothman, A. I. (1992). The development of a physiotherapy national examination OSCE. In *The 5th Ottawa Conference: Approaches to the assessment of clinical competence* (pp. 289-294). Ottawa, Canada.

Poldre, P., Smee, S. M., Reznick, R. K., Blackmore, D. E., Birtwhistle, R., Blouin, D., Chalmers, A., Galway, B., Hodges, B., MacFadyen, J., & Spady, D. (1999). The experience of thousands: The post-examination OSCE station review process of the Medical Council of Canada. In D. E. Melnick (Ed.), *Evolving assessment: Protecting the human dimension (CD-ROM)* Philadelphia: National Board of Medical Examiners.

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

Reilly, R. R., Henry, S., & Smither, J. W. (1990). An examination of the effects of using behavior checklists on the construct validity of assessment center dimensions. *Personnel Psychology*, 43, 71-84.

Resnick, L. B. & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. Gifford & M. C. O'Connor (Eds.), *Changing Assessments: Alternative Views of Aptitude, Achievement and Instruction* (pp. 37-75). Boston: Kluwer Academic.

Reznick, R. K., Blackmore, D. E., Cohen, R., Baumber, J. S., Rothman, A. I., Smee, S. M., Chalmers, A., Poldre, P., Birtwhistle, R., Walsh, P., Spady, D., & Bérard, M. J. (1993). An objective structured clinical examination for the licentiate of the Medical Council of Canada: From research to reality. *Academic Medicine*, S68, 4-6.

Reznick, R. K., Blackmore, D. E., Dauphinee, W. D., Smee, S. M., & Rothman, A. I. (1997). An OSCE for licensure: The Canadian experience. In A. J. J. A. Scherpbier, C. P. M. van der Vleuten, J. J. Rethans, & A. F. W. van der Steeg (Eds.), *Advances in Medical Education* (pp. 458-461). Dordrecht: Kluwer Academic Publishers.

Reznick, R. K., Blackmore, D. E., Dauphinée, W. D., Rothman, A. I., & Smee, S. M. (1996). Large-scale high-stakes testing with an OSCE: Report from the Medical Council of Canada. *Academic Medicine*, S71, 19-21.

Reznick, R. K., Regehr, G., Yee, G., Rothman, A. I., Blackmore, D. E., & Dauphinee, W. D. (1998). Process-rating forms versus task-specific checklists in an OSCE for medical licensure. *Academic Medicine*, 73, 97-99.

Sawhill, A. J., Dillon, G. F., Ripkey, D. R., Hawkins, R. E., & Swanson, D. B. (2003). The impact of postgraduate training and timing on USMLE Step 3 performance. *Academic Medicine*, 78, S10-S12.

Schmidt, H. G., Norman, G. R., & Boshuizen, P. A. (1990). A cognitive perspective on medical expertise: Theory and implications. *Academic Medicine*, 65, 611-621.

Schuwirth, L. W. T. & van der Vleuten, C. P. M. (2004). Changing education, changing assessment, changing research? *Medical Education*, 38, 805-812.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher*, 29, 4-14.

Snow, R. E. & Lohman, D. F. (1993). Cognitive psychology, new test design, and new test theory: An introduction. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 1-17). Hillsdale, NJ: Lawrence Erlbaum Associates.

Stake, R. (1999). The goods on American education. *Phi Delta Kappan*, May, 668-672.

Stemler, S. E. (2004). *A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. Practical Assessment, Research & Evaluation* Retrieved 3-10-2004 from http://PAREonlilne.net/getvn.asp?v=9&n=4

Sternberg, R. J. (1998). Abilities are forms of developing expertise. *Educational Researcher*, 27, 11-20.

Stiggins, R. J. (1991). Facing the challenges of a new era of educational assessment. *Applied Measurement in Education*, 4, 263-273.

Stiggins, R. J. (1994). Classroom perspectives on standardized testing. In *Student-centered classroom assessment* (pp. 329-359). New York: Merrill of Macmillan College Publishing.

Streiner, D. L. (1985). Global rating scales. In V. R. Neufeld & G. R. Norman (Eds.), *Assessing Clinical Competence* (pp. 119-141). New York: Springer Publishing.

Supovitz, J. A. & Brennan, R. L. (1997). Mirror, mirror on the wall, which is the fairest test of all? An examination of the equitability of portfolio assessment relative to standardized tests. *Harvard Educational Review*, 67, 472-502.

Swanson, D. B. & Norcini, J. J. (1989). Factors influencing reproducibility of tests using standardized patients. *Teaching and Learning in Medicine* 158-166.

Swanson, D. B., Norman, G. R., & Linn, R. L. (1995). Performance-based assessment: Lessons from the health professions. *Educational Researcher*, 24, 5-11, 35.

Tabachnick, B. G. & Fidell, L. S. (1996). *Using multivariate statistics*. (3rd ed.) New York: Harper Collins College Publishers.

van der Vleuten, C. P. M. (1996). The assessment of professional competence: Developments, research and practical implications. *Advances in Health Sciences Education: Theory and Practice*, 1, 41-46.

van der Vleuten, C. P. M., Norman, G. R., & De Graffe, E. (1991). Pitfalls in pursuit of objectivity: Issues of reliability. *Medical Education*, 25, 110-118.

van der Vleuten, C. P. M. & Swanson, D. B. (1990). Assessment of clinical skills with standardized patients: State of the art. *Teaching and Learning in Medicine*, 2, 58-76.

van Luijk, S. J. & van der Vleuten, C. P. M. (1992). A comparison of checklists and rating scales in performance-based testing. In I. R. Hart, R. M. Harden, & J. Des Marchais (Eds.), *Current Developments in Assessing Clinical Competence* (pp. 357-362). Montreal: Can-Heal Publications Inc.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan*, 70, 703-713.

Wilson, R. J. (1999). Aspects of validity in large-scale programs of student assessment. *The Alberta Journal of Educational Research*, 45, 333-343.

Wolfe, E. W. & Gitomer, D. H. (2001). The Influence of Changes in Assessment Design on the Psychometric Quality of Scores. *Applied Measurement in Education*, 14, 91-107.

Appendixes

INSTRUCTIONS TO THE EXAMINER: 10-minute communication station

AS THE CANDIDATE ENTERS THE ROOM, you should:

1. receive a bar code label from the CANDIDATE;
2. place the label in the box marked "CANDIDATE" on this page;
3. tell the candidate to proceed if they do not start right away;
4. candidates may review the Candidate's Instructions at any time.

WHILE THE CANDIDATE IS IN THE ROOM, you should:

1. closely observe the candidate-patient interaction;
2. fill in the appropriate bubbles on the checklist;
3. **DO NOT** interfere with the patient or the candidate except as indicated by the checklist or if the candidate is harming the patient;
4. if a candidate tries to leave the room prematurely, remind the candidate that they must wait for the buzzer that signals the END of the station. The candidate must wait quietly if they feel they have completed the station.

Note: The two halves of this sheet are coded with an unique serial number. It is very important that you do not mix sheets between candidates. Please do not separate sheets.

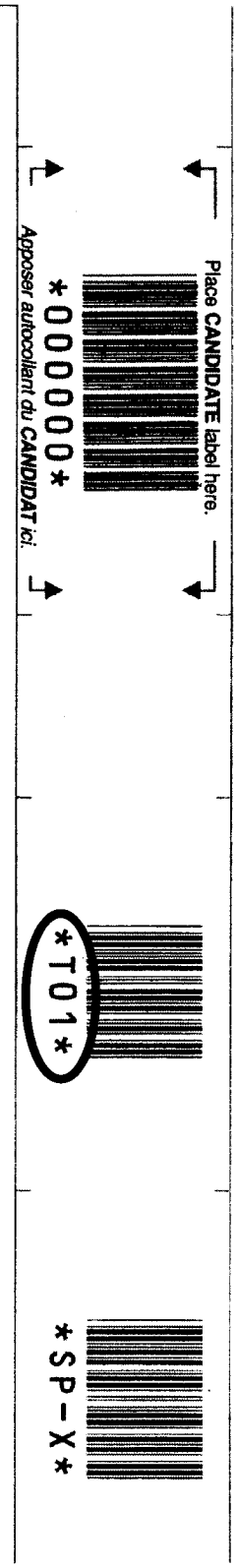
AFTER THE CANDIDATE HAS LEFT THE ROOM (and before the next candidate enters), you should quickly:

1. ensure that the appropriate bubbles have been filled in;
2. **COMPLETE** the global assessment questions at the bottom of the checklist;
3. move on to the Communication Checklist:
   - fill in one bubble for each category;
   - **answer the global assessment question** at the bottom of the communication checklist;
4. check to make sure the label is in place;
5. sign your name in the signature box (See the "X" on this page).

PURPOSE OF STATION

To take a history to elicit features of depression, to inquire into suicidal risk and to recognize importance of anniversary in presentation of depression.

SCORING GUIDELINES: N/A



(Front Page)

Done / completed satisfactorily	Done / completed satisfactorily
01AT/BT 1. **Elicits history of present complaint** (S) -sleep (sleeps more; sleeps more heavily) (S) -weight loss (yes, 5 kg/10 lb in 3 weeks) (S) -appetite (has to force down food) (S) -change in bowel habits (constipated) (S) -energy (decreased) (S) -activity (decreased) (S) -sexual drive (decreased) 2. **Elicits past medical history** (S) -medical history (none) (S) -psychiatric history (none) (S) -familial psychiatric history (none) 3. **Elicits associated symptoms** (S) -mood (irritable, moody, feels quite down) (S) -alcohol use (none) (S) -prescription drug use (none) (S) -non-prescription drug use (none) (S) -asks about suicide thoughts (none) 4. **Elicits further history** (S) -time of father's death (13 months ago) (S) -relationship with father (okay, I guess) (S) -relationship with mother and siblings (good) (S) -friends (no time for friends anymore) (S) -work situation (stressful)	

Did the candidate respond satisfactorily to the needs/problem(s) presented by this patient?

SATISFACTORY	UNSATISFACTORY
○ Borderline	○ Borderline
○ Good	○ Poor
○ Excellent	○ Inferior

If UNSATISFACTORY, please specify why:

(For items 4-6, please explain below)

(1) Inadequate medical knowledge and/or provided misinformation
(2) Could not focus in on this patient's problem
(3) Demonstrated poor communication and/or interpersonal skills
(4) Actions taken may harm this patient
(5) Actions taken may be imminently dangerous to this patient
(6) Other

(Back Page)

INSTRUCTIONS TO THE EXAMINER: 10-minute communication station with oral question(s)

AS THE CANDIDATE ENTERS THE ROOM, you should:

1. receive a bar code label from the CANDIDATE;
2. place the label in the box marked "CANDIDATE" on this page;
3. tell the candidate to proceed if they do not start right away;
4. candidates may review the Candidate's Instructions at any time.

WHILE THE CANDIDATE IS IN THE ROOM, you should:

1. closely observe the candidate-patient interaction;
2. fill in the appropriate bubbles on the checklist;
3. **DO NOT** interfere with the patient or the candidate except as indicated by the checklist or if the candidate is harming the patient;
4. if a candidate tries to leave the room prematurely, remind the candidate that they must wait for the buzzer that signals the END of the station. The candidate must wait quietly if they feel they have completed the station;
5. stop the candidate at the sound of the warning buzzer by telling them that you have questions that you would like to ask. Proceed with the question and mark their answers accordingly. You have one minute to administer the oral questions.

Note: The two halves of this sheet are coded with an unique serial number. It is very important that you do not mix sheets between candidates. Please do not separate sheets.

AFTER THE CANDIDATE HAS LEFT THE ROOM (and before the next candidate enters), you should quickly:

1. ensure that the appropriate bubbles have been filled in;
2. **COMPLETE** the global assessment questions at the bottom of the checklist;
3. move on to the Communication Checklist:
   - fill in one bubble for each category;
   - **answer the global assessment question** at the bottom of the communication checklist;
4. check to make sure the label is in place;
5. sign your name in the signature box (See the "X" on this page).

PURPOSE OF STATION

To elicit a history and perform a mental status examination of a patient in hospital who presents with delirium, recognizing alcohol withdrawal as the likely cause. Also, to determine the patient's capacity to consent to treatment by ascertaining her comprehension of the seriousness of her problem.

SCORING GUIDELINES:

Oral questions: Please review the answer key prior to exam start to ensure effective scoring.

Place CANDIDATE label here.
Apposer autocollant du CANDIDAT ici.
000000
T07
SP-X

(Front Page)

Done / completed satisfactorily

07AT/BT

1. **Elicits history of chief complaint**
   - -history of hallucinations
   - -still has hallucinations (no)
   - -description of hallucinations (very real)
   - -previous history of similar symptoms (none)
   - -past medical history (prior admission for liver problem)
   - -history of medication taken at home (none)
   - -past psychiatric history (none)

2. **Elicits history of alcohol intake**
   - -personal drinking history (had some problems before)
   - -recent drinking (began drinking again a few weeks ago)
   - -what has she been drinking (gin only)
   - -how much (26 to 40 oz/day)
   - -when did she drink last (just prior to admission)

3. **Elicits history of associated symptoms**
   - -fever (no)
   - -abdominal pain (hurts where the incision is)
   - -tremors (very shaky in the last 2 days)

4. **Performs mental status examination of orientation to time and day/date**
   - -time (does not know)
   - -place (at the hospital)
   - -person (can answer)
   - -immediate recall [e.g., repeat numbers or words] (poor)
   - -short term memory or new learning (recalls 1 of 3 objects after 3-5 minutes)
   - -remote memory (good)
   - -attention and concentration [e.g.; serial 3's or 7's or spell a word backwards] (poor)

5. **Explores patient's ability to understand and appreciate the nature of delirium/acute withdrawal** (denies seriousness)

6. **Provides information on:**
   - -the nature of the treatment [e.g., gradual withdrawal with medications, staying in hospital, further investigations]
   - -the risks of leaving the withdrawal/delirium untreated [e.g., unstable mental state, risk of death]

Done / completed satisfactorily

ORAL QUESTIONS (ONE MINUTE):

Questions to be administered ORALLY by the examiner. Note: You may only repeat the question once. If an answer is not forthcoming, then score "No answer or other answer given."

Q1. Do you think she's competent to decide to leave the hospital?
- -Yes
- -No

Q2. If this patient refuses to take her medication to prevent further hallucinations or worsening of her condition, what do you do?
[Examiner: Indicate the SINGLE item that best represents the candidate's response].
- -Obtain consent from her capable / competent next of kin (husband) or previously appointed substitute decider (e.g., a person designated by the patient in a Power of Attorney for personal care decisions)
- -Inform the patient that she is currently not capable of making treatment decisions and that they will proceed with treatment without her consent as this is considered a medical emergency, while trying to obtain consent from her capable/ competent next of kin (husband).
- -Inform the patient that they find her not capable of making treatment decisions at this time and give intramuscular medication if required urgently. Delirium is considered an emergency situation.
- -Document that the patient is incapable of making treatment decisions, using appropriate mental health legislated requirement for documentation (forms). Try to obtain consent from capable next of kin and/or proceed with treatment as this is an emergency.
- -Request consultation with a psychiatrist.
- -Accepts the patient's wishes.
- -No answer or other answer given.

Did the candidate respond satisfactorily to the needs/problem(s) presented by this patient?

SATISFACTORY	UNSATISFACTORY
○ Borderline	○ Borderline
○ Good	○ Poor
○ Excellent	○ Inferior

If UNSATISFACTORY, please specify why:
(For items 4-6, please explain below)

1. Inadequate medical knowledge and/or provided misinformation
2. Could not focus in on this patient's problem
3. Demonstrated poor communication and/or interpersonal skills
4. Actions taken may harm this patient
5. Actions taken may be imminently dangerous to this patient
6. Other

(Back Page)

Appendix B: Patient Interaction Rating Scale (PIRS)

EXAMINERS: Score the candidate on each of the items below by **filling** in the appropriate number. [Please note that the descriptors for each item are intended as helpful examples to use as guidelines for assessing candidate performance. The descriptors are not intended to be all-inclusive definitions.] In addition, answer the global assessment question at the bottom of the page.

1. Initiation of Interview

1	2	3	4	5	6
Lack of introduction	Minimal acknowledgement of patient	Borderline unsatisfactory, acknowledges patient, introduces self	Borderline satisfactory, acknowledges patient, introduces self	Acknowledges patient, moderately at ease & attentive	Attentive to patient, introduces self, at ease, personable

2. Listening Skills

1	2	3	4	5	6
Interrupts inappropriately, ignores patient's answers	Impatient	Borderline unsatisfactory, somewhat attentive	Borderline satisfactory, somewhat attentive	Attentive to patient's answers	Consistently attentive to answers & concerns

3. Questioning Skills

1	2	3	4	5	6
Awkward, exclusive use of closed ended or leading questions, jargon	Somewhat awkward, inappropriate terms, minimal use of open-ended questions	Borderline unsatisfactory, moderately at ease, appropriate language, uses different types of questions	Borderline satisfactory, moderately at ease, appropriate language, uses different types of questions	At ease, clear questions, appropriate use of open and closed ended questions	Confident and skilful questioning

4. Organization of Interview

1	2	3	4	5	6
Scattered, shot-gun approach	Minimally organized	Borderline unsatisfactory, flow is somewhat logical	Borderline satisfactory, flow is somewhat logical	Logical flow with sense of purpose	Purposeful, integrated handling of encounter

5. Closing

1	2	3	4	5	6
Abrupt	Acknowledges end of interview	Borderline unsatisfactory, attempts closure	Borderline satisfactory, attempts closure	Clear closure	Organized, thoughtful closure

6. Rapport with Person

1	2	3	4	5	6
Condescending, offensive, judgmental	Minimal courtesies only	Borderline unsatisfactory	Borderline satisfactory	Polite and interested	Warm, polite, empathic

7. Professional Behavior

1	2	3	4	5	6
Offensive or aggressive; frank exhibition of "unprofessional conduct"	Negative attitude to patient	Borderline unsatisfactory, does not truly instil confidence	Borderline satisfactory, manner inoffensive but does not necessarily instil confidence	Attempts professional manner with some success	Overall demeanour of a professional, caring, listens, communicates effectively

SATISFACTORY	**UNSATISFACTORY**
○ **Borderline**	○ **Borderline**
○ **Good**	○ **Poor**
○ **Excellent**	○ **Inferior**

» **Overall, do you feel that the candidate communicated/behaved with this patient in a satisfactory manner? If Unsatisfactory, please specify why:**

INSTRUCTIONS TO THE EXAMINER: 10-minute communication station

AS THE CANDIDATE ENTERS THE ROOM, you should:

1. receive a bar code label from the CANDIDATE;
2. place the label in the box marked "CANDIDATE" on this page.

WHILE THE CANDIDATE IS IN THE ROOM, you should:

1. closely observe the candidate-patient interaction;
2. fill in the appropriate bubbles on the rating scale;
3. **DO NOT** interfere with the patient or the candidate.

Note: The two halves of this sheet are coded with an unique serial number. It is very important that you do not mix sheets between candidates. Please do not separate sheets.

AFTER THE CANDIDATE HAS LEFT THE ROOM (and before the next candidate enters), you should quickly:

1. ensure that the appropriate bubbles have been filled in;
2. **COMPLETE** the global assessment questions at the bottom of the page;
3. move on to the Communication Checklist:
   - fill in each of the bubbles in each category that applied to the candidate;
   - **answer the global assessment question** at the bottom of the page;
4. check to make sure the label is in place;
5. sign your name in the signature box (See the "X" on this page).

Place CANDIDATE label here.

Apposer autocollant du CANDIDAT ici.

000000

PURPOSE OF STATION

To take a history to elicit features of depression, to inquire into suicidal risk and to recognize importance of anniversary in presentation of depression.

CASE SUMMARY:

Lisa Green (24 years old) presents due to weight loss and stress from poor work performance, which is a concern as her income is critical to her family since her father was killed in a car accident. Her current symptoms started 3-4 weeks ago and include feeling weary and irritable, suffering from loss of energy, loss of appetite, weight loss and constipation. She is sleeping more but feels at her worst in the mornings. Her problems at work started out "a few months ago". Five months ago she ended her relationship with a boyfriend - "lost interest". Her father was killed 13 months ago. Lisa coped well at the time and went back to work soon after. She also helps with the housekeeping and with her siblings (15 and 17 years old). Her mother is now working part-time as a housekeeper and she has been talking more about the accident for the past 3-4 weeks. Lisa is depressed secondary to her bereavement but she is not suicidal.

R01

SP-X

(Front Page)

EXAMINERS: Score the candidate on each of the items below by **filling** in the appropriate number. [Please note that the descriptors for each item are intended as helpful examples to use as guidelines for assessing candidate performance. The descriptors are not intended to be all-inclusive definitions.] In addition, answer the global assessment question at the bottom of the page. The rating scale runs from:

Inferior	Poor	Borderline unsatisfactory	Borderline satisfactory	Good	Excellent

1. Elicits relevant history of patient's weight loss and poor work performance:

Absent or inappropriate history of chief complaint; e.g., only considers medical / weight loss issues or focuses on stress / work		Borderline unsatisfactory Elicits partial history of chief complaint that includes consideration of both medical and psychiatric diagnoses	Borderline satisfactory		Uses pertinent questions to focus on history of chief complaint; considers both medical and psychiatric diagnoses

2. Elicits factors associated with primary diagnosis of depression:

No assessment of degree of depression (e.g., risk of suicide) and / or ignores contributing factors; e.g., medications, drugs, alcohol; poor focus		Borderline unsatisfactory Partial assessment of degree of depression (e.g., risk of suicide) and / or contributing factors; e.g., medications, drugs, alcohol	Borderline satisfactory		Clear line of thought in assessing degree of depression (e.g., risk of suicide) and / or contributing factors; e.g., medications, drugs, alcohol

3. Assesses relevant past medical history:

No or inappropriate assessment of contribution of past medical and psychiatric history to current health		Borderline unsatisfactory Limited assessment of contribution of past medical and / or psychiatric history to current health	Borderline satisfactory		Appropriate assessment of contribution of past medical and / or psychiatric history to current health

4. Assess impact of depression on patient's work and personal life:

Lack of or negative consideration given to the impact of depression on patient's work and personal situation		Borderline unsatisfactory Limited consideration and empathy given to the impact of depression on patient's work and personal situation	Borderline satisfactory		Appropriate and empathetic consideration and empathy given to the impact of depression on patient's work and personal situation

5. Elicits relevant history regarding father's death:

Lack of or superficial attention given to father's death or date of death		Borderline unsatisfactory Some attention paid to father's death but does not seem to recognize significance of anniversary; limited empathy	Borderline satisfactory		Attention paid to father's death in an empathetic manner, significance of anniversary of his death to patient's depression is clearly recognized

6. Explores appropriately the patient's perception / comprehension of her problems:

Ignores, misses or is hostile to the patient's concern regarding her ability to do her job and meet family responsibilities		Borderline unsatisfactory Acknowledges patient's concerns about her job and family responsibilities but limited or absent response to her issues	Borderline satisfactory		Acknowledges and responds effectively to patient's concern regarding her ability to do her job and meet family responsibilities

SATISFACTORY	UNSATISFACTORY
◯ Borderline	◯ Borderline
◯ Good	◯ Poor
◯ Excellent	◯ Inferior

» Did the candidate respond satisfactorily to the needs/problem(s) presented by this patient? If UNsatisfactory, please specify why:

DO NOT write in margins

(Back Page)

INSTRUCTIONS TO THE EXAMINER: 10-minute communication station with oral question(s)

AS THE CANDIDATE ENTERS THE ROOM, you should:

1. receive a bar code label from the CANDIDATE;
2. place the label in the box marked "CANDIDATE" on this page.

WHILE THE CANDIDATE IS IN THE ROOM, you should:

1. closely observe the candidate-patient interaction;
2. fill in the appropriate bubbles on the rating scale;
3. DO NOT interfere with the patient or the candidate.

Note: The two halves of this sheet are coded with an unique serial number. It is very important that you do not mix sheets between candidates. Please do not separate sheets.

AFTER THE CANDIDATE HAS LEFT THE ROOM (and before the next candidate enters), you should quickly:

1. ensure that the appropriate bubbles have been filled in;
2. **COMPLETE** the global assessment questions at the bottom of the page;
3. move on to the Communication Checklist:
   - fill in each of the bubbles in each category that applied to the candidate;
   - **answer the global assessment question** at the bottom of the page;
4. check to make sure the label is in place;
5. sign your name in the signature box (See the "X" on this page).

Place **CANDIDATE** label here.

*Apposer autocollant du **CANDIDAT** ici.*

000000

PURPOSE OF STATION

To elicit a history and perform a mental status examination of a patient in hospital who presents with delirium, recognizing alcohol withdrawal as the likely cause. Also, to determine the patient's capacity to consent to treatment by ascertaining her comprehension of the seriousness of her problem.

CASE SUMMARY:

Sylvia Brenner (58 years old) had an uncomplicated hysterectomy for fibroids 3 days ago. Two days ago she began feeling shaky and nervous. Last night she was hallucinating and agitated. She has received Tylenol 3® and lorazepam (1 mg) at bedtime. She knows she was confused but the hallucinations are real to her - she was awake, she saw and heard people partying. Her PMH is unremarkable except for a bout of jaundice due to alcoholic hepatitis that required hospitalization 4 years ago. Her drinking had been erratic but heavy. She had had blackouts and been fired for absenteeism. Prior to that hospitalization she had stopped drinking and begun attending AA meetings, which she kept up until about 6 months ago. Her "worries" about this surgery led to drinking and just prior to admission she was consuming 750 ml - 1 litre of gin/day. She has had none since her admisssion. Sylvia is poorly oriented to time and place, unable to concentrate and has poor recall. She does not accept that she is seriously ill, she just wants to go home.

R07

SP-X

(Front Page)

EXAMINERS: Score the candidate on each of the items below by filling in the appropriate number. [Please note that the descriptors for each item are intended as helpful examples to use as guidelines for assessing candidate performance. The descriptors are not intended to be all-inclusive definitions.] In addition, answer the global assessment question at the bottom of the page. The rating scale runs from:

Inferior	Poor	Borderline unsatisfactory	Borderline satisfactory	Good	Excellent

1. Elicits relevant history of patient's delirium:

Absent or inappropriate history taking regarding delirium	Borderline unsatisfactory	Borderline satisfactory	Uses pertinent questions to focus on relevant history of delirium
	Elicits a partial but relevant history of the delirium		

2. Elicits relevant history of symptoms to determine causes of delirium:

Absent or irrelevant history related to establishing the cause of this patient's delirium; no clear line of thought	Borderline unsatisfactory	Borderline satisfactory	Considers multiple causes for this patient's delirium; e.g., asks about medications *and* post-operative infection
	Partial history relevant to establishing the cause of this patient's delirium; e.g.. asks about medications *or* post-operative infection		

3. Explores relevant past medical history:

Absent or irrelevant past medical or psychiatric history and / or takes inappropriately detailed history	Borderline unsatisfactory	Borderline satisfactory	Takes relevant past medical and psychiatric history with an appropriate amount of detail
	Takes partial past medical and / or psychiatric history with some degree of relevance and appropriateness		

4. Elicits relevant history of alcohol intake:

No consideration or assessment of alcohol withdrawal as cause of delirium	Borderline unsatisfactory	Borderline satisfactory	Appropriate and effective history establishing alcohol withdrawal as probable cause of delirium
	Some consideration given to alcohol withdrawal as cause of delirium		

5. Conducts focused assessment of mental status (e.g., orientation, recall, concentration):

No mental status assessment or inferior assessment; e.g., overly limited or unnecessarily detailed	Borderline unsatisfactory	Borderline satisfactory	Effective and relevant assessment of mental status
	Assesses mental status with some degree of skill and focus		

6. Assesses patient's perception / comprehension of problem (include their assessment of patient & their response to oral questions):

Ignores, misses or is hostile to the patient's limited understanding of seriousness of her problem and perceives her as competent to consent; e.g., accepts her wishes / decision.	Borderline unsatisfactory	Borderline satisfactory	Recognizes patient's limited understanding of her problem and her incompetence to make treatment decisions; e.g., wants consent of next of kin or substitute decision-maker, will treat her anyway while seeking consent as this is an emergency
	Recognizes patient's limited understanding of her problem but does not fully grasp the issue of competence and urgent need for treatment; e.g., refers to psychiatrist		

SATISFACTORY	UNSATISFACTORY
Borderline	Borderline
Good	Poor
Excellent	Inferior

» Did the candidate respond satisfactorily to the needs/problem(s) presented by this patient? If UNsatisfactory, please specify why:

DO NOT write in margins

(Back Page)

INSTRUCTIONS TO THE EXAMINER: 10-minute communication station

AS THE CANDIDATE ENTERS THE ROOM, you should:

1. receive a bar code label from the CANDIDATE;
2. place the label in the box marked "CANDIDATE" on this page.

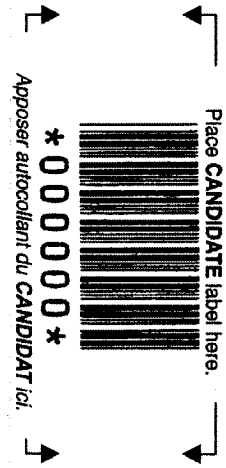
WHILE THE CANDIDATE IS IN THE ROOM, you should:

1. closely observe the candidate-patient interaction;
2. fill in the appropriate bubbles on the rating scale;
3. **DO NOT** interfere with the patient or the candidate.

Note: The two halves of this sheet are coded with an unique serial number. It is very important that you do not mix sheets between candidates. Please do not separate sheets.

AFTER THE CANDIDATE HAS LEFT THE ROOM (and before the next candidate enters), you should quickly:

1. ensure that the appropriate bubbles have been filled in;
2. **COMPLETE** the global assessment questions at the bottom of the page;
3. move on to the Communication Checklist:
   - fill in each of the bubbles in each category that applied to the candidate;
   - **answer the global assessment question** at the bottom of the page;
4. check to make sure the label is in place;
5. sign your name in the signature box (See the "X" on this page).

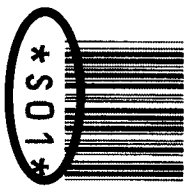


PURPOSE OF STATION

To take a history to elicit features of depression, to inquire into suicidal risk and to recognize importance of anniversary in presentation of depression.

CASE SUMMARY:

Lisa Green (24 years old) presents due to weight loss and stress from poor work performance, which is a concern as her income is critical to her family since her father was killed in a car accident. Her current symptoms started 3-4 weeks ago and include feeling weary and irritable, suffering from loss of energy, loss of appetite, weight loss and constipation. She is sleeping more but feels at her worst in the mornings. Her problems at work started out "a few months ago". Five months ago she ended her relationship with a boyfriend - "lost interest". Her father was killed 13 months ago. Lisa coped well at the time and went back to work soon after. She also helps with the housekeeping and with her siblings (15 and 17 years old). Her mother is now working part-time as a housekeeper and she has been talking more about the accident for the past 3-4 weeks. Lisa is depressed secondary to her bereavement but she is not suicidal.



(Depression – Front page)

INSTRUCTIONS TO THE EXAMINER: 10-minute communication station with oral question(s)

AS THE CANDIDATE ENTERS THE ROOM, you should:

1. receive a bar code label from the CANDIDATE;
2. place the label in the box marked "CANDIDATE" on this page.

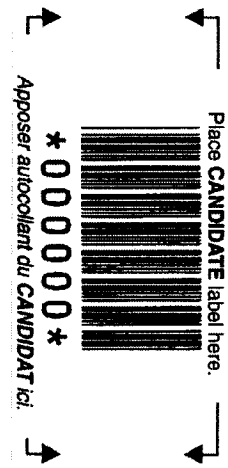
WHILE THE CANDIDATE IS IN THE ROOM, you should:

1. closely observe the candidate-patient interaction;
2. fill in the appropriate bubbles on the rating scale;
3. DO NOT interfere with the patient or the candidate.

Note: The two halves of this sheet are coded with an unique serial number. It is very important that you do not mix sheets between candidates. Please do not separate sheets.

AFTER THE CANDIDATE HAS LEFT THE ROOM (and before the next candidate enters), you should quickly:

1. ensure that the appropriate bubbles have been filled in;
2. **COMPLETE** the global assessment questions at the bottom of the page;
3. move on to the Communication Checklist:
   - fill in each of the bubbles in each category that applied to the candidate;
   - **answer the global assessment question** at the bottom of the page;
4. check to make sure the label is in place;
5. sign your name in the signature box (See the "X" on this page).

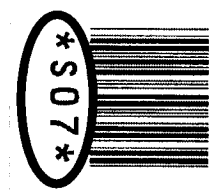


PURPOSE OF STATION

To elicit a history and perform a mental status examination of a patient in hospital who presents with delirium, recognizing alcohol withdrawal as the likely cause. Also, to determine the patient's capacity to consent to treatment by ascertaining her comprehension of the seriousness of her problem.

CASE SUMMARY:

Sylvia Brenner (58 years old) had an uncomplicated hysterectomy for fibroids 3 days ago. Two days ago she began feeling shaky and nervous. Last night she was hallucinating and agitated. She has received Tylenol 3® and lorazepam (1 mg) at bedtime. She knows she was confused but the hallucinations are real to her - she was awake, she saw and heard people partying. Her PMH is unremarkable except for a bout of jaundice due to alcoholic hepatitis that required hospitalization 4 years ago. Her drinking had been erratic but heavy. She had had blackouts and been fired for absenteeism. Prior to that hospitalization she had stopped drinking and begun attending AA meetings, which she kept up until about 6 months ago. Her "worries" about this surgery led to drinking and just prior to admission she was consuming 750 ml - 1 litre of gin/day. She has had none since her admisssion. Sylvia is poorly oriented to time and place, unable to concentrate and has poor recall. She does not accept that she is seriously ill, she just wants to go home.



(Delirium – Front Page)

EXAMINERS: Score the candidate on each of the items below by **filling** in the appropriate number. [Please note that the descriptors for each item are intended as helpful examples to use as guidelines for assessing candidate performance. The descriptors are not intended to be all-inclusive definitions.] In addition, answer the global assessment question at the bottom of the page. The rating scale runs from:

Inferior **Poor** **Borderline unsatisfactory** **Borderline satisfactory** **Good** **Excellent**

1. Elicits relevant history of presenting problem:

Absent or superficial effort to elicit a description of the presenting problem	Borderline unsatisfactory / Borderline satisfactory: Elicits a partial description of the presenting problem	Uses pertinent questions to focus on key aspects of the presenting problem and to elicit description of the presenting problem

2. Elicits relevant history of associated symptoms:

Absent or irrelevant history of associated symptoms, given the presenting problem; no focus	Borderline unsatisfactory / Borderline satisfactory: Partial history of relevant associated symptoms, given the presenting problem	Focused, comprehensive history of relevant associated symptoms, given the presenting problem

3. Elicits further history relevant to differential diagnoses (e.g., ROS):

Absent or irrelevant review of systems and history of contributing factors; no clear line of thought	Borderline unsatisfactory / Borderline satisfactory: Limited review of systems and / or history of contributing factors; line of thought may be unclear	Relevant review of systems and history of contributing factors, shows a clear line of thought

4. Elicits history of medications (e.g., prescriptions, over-the-counter and illicit drugs):

Ignores or largely misses history of medications and other drugs	Borderline unsatisfactory / Borderline satisfactory: Attempts to take history of medications but may be too focused or of limited relevance	Takes skilful history of all relevant medications and drugs

5. Elicits relevant past medical history (given nature and urgency of the presenting problem):

No past medical history and / or asks for inappropriate detail	Borderline unsatisfactory / Borderline satisfactory: Some relevant past medical history with a certain degree of appropriate detail	Relevant and appropriate past medical history

6. Elicits relevant psycho-social history (given nature and urgency of the presenting problem):

Absent or inappropriate consideration given to the patient's situation	Borderline unsatisfactory / Borderline satisfactory: Some consideration given to the patient's situation	Appropriate consideration given to the patient's situation

7. Explores the patient's perception / comprehension of the presenting problem:

Ignores, misses or is hostile to the patient's perceptions and has a poor grasp of the issue(s) and / or the patient's affect	Borderline unsatisfactory / Borderline satisfactory: Expresses some awareness of patient's perceptions and shows some grasp of the issue(s) and / or affect; may attempt to address the concerns but with limited effectiveness	Expresses awareness of patient's perceptions and has a good grasp of issue(s) and / or affect and responds effectively

SATISFACTORY	**UNSATISFACTORY**
Borderline	**Borderline**
Good	**Poor**
Excellent	**Inferior**

» Did the candidate respond satisfactorily to the needs/problem(s) presented by this patient? If UNsatisfactory, please specify why:

DO NOT write in margins

(Skill-specific Rating Scale - Back Page)

Appendix E: Feedback Form for Assessors

Comparing Scoring Instruments for the Performance Assessment of Professional Competencies - *Feedback Form*

Name: _____

Faculty Position (if applicable): _____

Specialty:_____ DOB: _____/___/_____ Male
Month/Day/Year Female

MCC ID # (see payment sheet): _____ ____

1. How many times have you been an examiner for the MCCQE Part II prior to this administration?

Never One time Two times Three times Four times Five or more times

2. Have you been an examiner for other assessments? If yes, please specify:

RCPSC or CFPC examinations

Undergraduate medical school assessments

Postgraduate assessments

Other – please specify:_____

3. Did you have sufficient time to score?

1	2	3	4	5
Insufficient	Not quite Sufficient	Barely sufficient	Sufficient	More than sufficient

4. If you used the rating scale score sheet (two rating scales – no checklist) please answer the following:

a. Have you used similar rating scales for other assessments?

Never One time Two times Three times Four times Five or more times

b. Did the anchors for the items on the content-based rating scale (left page) provide sufficient information for you to score accurately?

1	2	3	4	5
Insufficient	Not quite Sufficient	Barely sufficient	Sufficient	More than sufficient

c. Did the items on the content-based rating scale (left page) comprehensively capture the candidate performances you observed?

1	2	3	4	5
Did not capture performance at all	Minimal capture of performance	Partially captured performance	Acceptable capture of performance	Comprehensive capture of performance

d. Do you have a preference for using a rating scale (left page) over a checklist?

1	2	3	4	5
Do not wish to use rating scale	Rather not use rating scale	No preference	Rather use rating scale	Much prefer rating scale

Comments:

_____

Appendix F: Consent Form for Assessors

CONSENT FORM for RESEARCH INITIATIVES
PHYSICIAN EXAMINERS
MCCQE Part II Fall 2003

The Medical of Canada is continuously assessing and revising the content and format of the Qualifying Examinations to ensure that these measures of clinical knowledge, skills and judgment are as valid and reliable as is feasible. Periodically this requires collection of data during the examination so that new or revised scoring instruments, question formats, and quality assurance processes can be evaluated.

Based on data showing strong performance of the Medical Council of Canada Qualifying Examination Part II in its 14-station format (1998-2003), the format of the examination will be altered as of the Fall 2003 administration to promote its further development. In addition, two research initiatives will be implemented. The details of these changes are listed here.

1. The examination will be comprised of 8 ten-minute stations and 6 couplet or five+ five stations. Previously, the exam has been comprised of an equal number of both types of station.

2. One ten-minute station and one couplet station will be pilots. As a result, pass-fail decisions will be based on 12 stations, not 14.

3. In some stations, there will be two observers, scoring independently of each other. One will be the assigned examiner for the station, and that physician's scores (based on the existing checklist and rating scale for the station) will be the basis for decision-making. The second observer will be from one or the other of the two projects summarized below:

a. Study One – Trained Assessors: An internal research initiative lead by *Dr. Sue Humphrey, Secretary of the OSCE Test Committee (613 737-8899 ext 78979)* is assessing the use of non-physician assessors to score in selected history-taking stations. The intent is to assess whether the demand for physicians can be reduced so that physician resources can be focused on those stations where their judgment is most critical to assessing the clinical ability of the candidates. Therefore, in some centers, in two of the short history-taking stations, a second assessor will observe and score candidate performance using the same checklist as the physician examiner. The goal of the pilot study is to establish selection criteria for trained assessors and to develop a suitable training protocol for them. Further evaluation of this approach will occur in subsequent administrations of the examination.

b. Study Two – Rating Scales: The second initiative is a doctoral thesis project from the University of Ottawa lead by Sydney Smee (613 521-6603), supervised by Dr. Dany Laveault, Faculty of Education, and funded by the Medical Council of Canada. The study is assessing the validity of two different forms of rating scales as alternate formats to the existing checklists with the intent of better capturing the expert judgment of the observing physician. For this project, a second physician assessor will be present in two of the ten-minute history-taking stations.

Proportionally, only a small group of physician examiners will be impacted by either of these projects. Those who are assigned to one of the selected stations, either as the acting examiner or as an assessor using a rating scale, will be asked to respect the research protocols.

These are not onerous and it is hoped that many of you will agree to be available for both these studies. (The protocol is summarized below.) The researcher, the University of Ottawa and the Medical Council of Canada will be scrupulously careful with respect to ensuring the confidentiality of test taker and examiner data. Only group data and anonymous quotes will be reported.

The results of both studies will be submitted for presentation at medical education meetings and for publication. Summary conclusions will be made available to participants by request to the manager of the MCCQE Part II. Your consideration of and support for this work is appreciated. Thank you.

PROTOCOL

As a participating examiner you will be expected to follow a simple protocol:

1. Score candidate performance independently, as you would for any other case, without indicating verbally or non-verbally your judgments to the other examiner/assessor.

2. Complete a short feedback form following the examination to indicate your judgment of the alternate scoring instrument (e.g., validity of content, ease of use, etc.) and to provide data regarding your background as an examiner (e.g., specialty, experience as an examiner for other assessments, previous experience with rating scales for assessment, etc.)

Information from all participants will be stored securely and all reports will be based on group data. Participant confidentiality will be respected throughout the research process. If, at any time, you wish to withdraw from the study, you may do so without penalty. All examiners will receive the honorarium regardless of whether they participate in either of the research initiatives.

CONSENT

I, _____, agree to be available as either a physician examiner or as an assessor for the research cases for the Fall 2003 MCCQE Part II as they have been described on this form and in the written Notice of Research Initiatives for Physician Examiners and as confirmed below – PLEASE CHECK ONE OR TWO OF THE BOXES BELOW TO INDICATE YOUR WISHES:

Yes, I DO agree to be available as an examiner for Study One – Trained Assessors

Yes, I DO agree to be available as an examiner or as an assessor for Study Two– Rating Scales

No, I do NOT agree to be available as an examiner for either study.

_____

Signature Date

_____

Witness - Signature Date

NOTE: One copy of this form is for your records. Please return the second copy to your examination center as soon as is convenient. (Local fax information to be inserted here.)

Appendix G: PCA Analysis

PCA is an approach to data reduction that transforms the original variables into uncorrelated variables that explain decreasing degrees of the variance in the data. The intent is to reduce the dimensionality of the data in a meaningful manner. Interpretation of the results depends on data being appropriate for the analysis. One indicator is the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy. This statistic is the ratio of the sum of the squared correlations to the sum of the squared correlations plus the sum of the squared partial correlations and values range from zero to one. Higher values indicate less partial correlations, meaning less error. The recommended minimum standard is 0.6 (Tabachnick & Fidell, 1996). Since these studies were only using the PCA to determine dimensionality, the data only needed to meet minimum standards.

Study One.

For CRS-1, KMO=.86 and the PCA extracted one component, based on eigenvalues >1. The first component explained 60% of variance and had an eigenvalue of 3.6, indicating a unidimensional scale. No other eigenvalue exceeded one.

CRS-1 combined with the PIRS had KMO=.948, one component based on eigenvalues>1 and the first component explained 59% of variance with an eigenvalue of 7.12.

For CL-1, KMO=0.65 with five of the 20 items exhibiting uneven splits of 90:10 or higher. PCA results for binary data are hard to interpret and made mores so by extreme splits like these ones. However, it can still provide an adequate overall sense of structure (McDonald, 1999) and so the CL-1 results are reported. Eight components had eigenvalues>1 and together explained 53% of the variance. The first component explained only 11.4% of the variance (eigenvalue=2.3).

For CL-1 combined with PIRS, KMO=.867 and there were nine components based on eigenvalues>1 that explained 57% of the variance. The first component

explained 19% and had an eigenvalue=5.0. Of interest, the first component was largely defined by the PIRS items.

For CRS-2, KMO=0.82 and one component was extracted, based on eigenvalues>1. The first component explained 50.4% of variance and had an eigenvalue of 3.0. No other eigenvalue exceeded one.

For CRS-2 combined with PIRS, KMO =.921 and two components were extracted based on eigenvalues>1. The two components explained 60% of the variance, with the eigenvalue=5.8 for the first component. All items contributed to the first component and 3 items from CRS-2 largely defined the second component.

For CL-2, the KMO value was not generated because the analysis of the correlation matrix did not allow a value to be calculated, suggesting that there might be multicollinearity or a singularity within the data. However, multicollinearity does not interfere with PCA (Tabachnick & Fidell, 1996) and tolerance values<1 indicated there was no singularity. Given the limitations of the analyses, only broad conclusions could be drawn but they were all that was required for this discussion. Ten components had eigenvalues>1 and together explained 64% of the variance. The first component explained 12.6% of the variance (eigenvalue=3.3).

For CL-2 combined with PIRS, 11 components were extracted based on eigenvalues>1 and they explained 11% of the variance. The first component explained 16% of the variance and had an eigenvalue =5.1. The PIRS items were the defining elements of the first component.

Study Two.

For the skill-specific rating scale, KMO=0.88 and 0.86 for Depression and Delirium respectively. One component was extracted for each case, based on eigenvalues>1. With Depression, the first component explained 52.7% of variance and had an eigenvalue of 3.7. No other eigenvalue exceeded one. With Delirium, the first

component explained 54.5% of the variance and had an eigenvalue of 3.8. No other eigenvalue exceeded one.

For the skill-specific rating scale combined with PIRS, KMO=.949 and .880 for Depression and Delirium respectively. For each case, two components were extracted, explaining 64% and 58% of the variance respectively. The first component for Depression explained 55% of the variance and had an eigenvalue=7.1. The first component included all the items; the second was defined by three of the items from the skill-specific rating scale. The first component for Delirium explained 38% of the variance and had an eigenvalue=4.9%. As with Depression the first component included all the items; but the second component was defined by the skill-specific rating scale items.

For CL-1, KMO=0.58. Eight components had eigenvalues>1 (as with the Study One analysis) and together explained 51% of the variance. The first component explained only 10.2% of the variance (eigenvalue=2.0).

For CL-1 combined with the PIRS, KMO=.859, 10 components were extracted explaining 60% of the variance. The first component explained 19% of the variance and had an eigenvalue=4.9. Items from the PIRS largely defined the first component.

For CL-2, KMO=0.86. Nine components had eigenvalues>1 (one less than found in the Study One analysis) and together explained 61% of the variance. The first component explained only 13.2% of the variance (eigenvalue=3.4).

For CL-2 combined with the PIRS there was no KMO value but 10 components were extracted, explaining 61% of the variance. The first component explained 15% of the variance and had an eigenvalue of 4.9. As with CL-1 the first component was largely defined by the PIRS items.

Appendix H: Rater Agreement

Study One - Depression

	PIRS Rating	PIRS PF	Significance
Pearson's R	0.486	0.339	0.000
Spearman Correlation	0.452	0.339	0.000
Kappa	0.193	0.338	0.000
Agreement[a]	93.2%	84.7%	
N of Cases		778	

Study One - Delirium

	PIRS Rating	PIRS PF	Significance
Pearson's R	0.391	0.284	0.000
Spearman Correlation	0.38	0.284	0.000
Kappa	0.158	0.284	0.000
Agreement[a]	87.0%	76.0%	
N of Cases		776	

Study Two - Depression

	PIRS Rating	PIRS PF	Significance
Pearson's R	0.536	0.398	0.000
Spearman Correlation	0.535	0.398	0.000
Kappa	0.254	0.397	0.000
Agreement[a]	93.8%	86.1%	
N of Cases		741	

Study Two - Delirium

	PIRS Rating	PIRS PF	Significance
Pearson's R	0.369	0.270	0.000
Spearman Correlation	0.403	0.270	0.000
Kappa	0.157	0.269	0.000
Agreement[a]	90.1%	76.3%	
N of Cases		670	

[a]*Agreement on rating means raters are within one point of each other on a six-point scale.*

Appendix I: Interpreting Logistic Regression Output

Goodness of fit.

With regression models fit refers to the degree to which the independent or predictor variables explain the variance in the data, which leads to how accurately they predict the observed values of the dependent variable. If the predictor variables do not predict the dependent variable significantly better than it would be predicted by chance then the model is not useful for explaining the dependent variable or in this research, for comparing instruments.

Two indicators of how well the model fits the data for logistic regression are the full model chi square (χ^2) and the Hosmer and Lemeshow statistics. If significant, the χ^2 statistic indicates the model-based predictions are better than would be predicted by chance. When the Hosmer and Lemeshow statistic is not significant, it indicates no significant difference in predictions between the observed model and an ideal model. This is a different approach to indicating the same outcome, namely that using the model improves prediction of the dependent variable. There is no agreement on one measure of model fit being the "best". Instead, several measures may be examined together to draw some conclusion about the fit and the nature of the relationship between the predictors and the dependent variables. For this research, both the χ^2 statistic and Hosmer and Lemeshow statistic were reported. They were interpreted in conjunction with measures of the degree of variance explained by the model.

Degree of variance explained by the model.

With a linear regression model, R^2 indicates the proportion of variance in the data explained by the model. With logistic regression, Nagelkerke R^2 and R_L^2 can be similarly interpreted. The Nagelkerke R^2 statistic indicates to what degree the predictors in the model improve the log likelihood (likelihood of producing the observed scores) relative to chance (Menard, 2001; Pampel, 2000). Menard advocates using a related measure, R_L^2,

which represents the degree to which the log likelihood is being maximized by the inclusion of the variables. Unlike Nagelkerke R^2 or other similar measures, R_L^2 is not dependent on sample size nor is it sensitive to the split in the dependent variable (e.g., the ratio of pass to fail). R_L^2 is simply the ratio of the χ^2 statistic for the full model to the sum of the χ^2 statistic for the full model and the χ^2 statistic for the block model. The values of both these statistics run from zero to one and the larger the value, the stronger the association. As ratios, they can also be read as the percentage of variance in the decisions explained by the model. While both values are reported in the tables, only R_L^2, the more conservative measure, was discussed.

Predicted versus Observed Pass-Fail.

A component of the results of the logistic regression is a classification table comparing predicted to observed pass-fail decisions. In order to compare the information in the classification tables across instruments, a summary measure (τ_p) was calculated for each one (Menard, 2001). The τ_p measure uses the proportional change in error as an indicator of predictive efficiency from the general form of

$$\text{Predictive efficiency} = \frac{(\text{expected errors without model}) - (\text{observed errors with model})}{(\text{expected errors without model})}$$

This measure is also analogous to R^2. As its value approaches one, it indicates that the model is increasingly accurate in classifying the pass-fail decisions. The significance of τ_p is indicated by the binomial d statistic, which is approximately normally distributed (Bulmer, 1979). With this distribution, $d > 3.0$ indicates that model is significantly more accurate in predicting pass-fail decisions than chance.

Appendix J: Observed correlations for Study One and Study Two instrument

Table 30 *Score correlations across instruments for two cases for Study One*

	CRS-1	CRS-1 PIRS	CL-1	CL-1 PIRS
CRS-1	.87			
CRS-1 PIRS	**.821**	.91		
CL-1	**.299**	.251	.55	
CL-1 PIRS	.421	**.478**	**.464**	.92
	CRS-2	**CRS-2 PIRS**	**CL-2**	**CL-2 PIRS**
CRS-2	.79			
CRS-2 PIRS	**.697**	.87		
CL-2	**.500**	.319	.64	
CL-2 PIRS	.368	**.429**	**.535**	.84

Note. Pearson correlations, all significant ($p < .01$; two-tailed)

Table 31 *Score correlations across instruments for two cases for Study Two*

	SRS-1	SRS-1 PIRS	CL-1	CL-1 PIRS
SRS -1	.84			
SRS-1 PIRS	**.740**	.91		
CL-1	**.356**	.269	.47	
CL-1 PIRS	.437	**.553**	**.395**	.92
	SRS-2	**SRS-2 PIRS**	**CL-2**	**CL-2PIRS**
SRS-2	.86			
SRS-2 PIRS	**.683**	.87		
CL-2	**.411**	.330	.60	
CL-2 PIRS	.434	**.434**	**.456**	.86

Note. Pearson correlations, all significant ($p < .01$, (two-tailed)

Appendix K: Alternate Expertise Models

Three variables, Knowledge, Skill and Experience, representing three aspects of expertise were selected to represent clinical competence for these studies. Using three variables made for a simplified model of a complex construct but it was rationalized that that these three could be sufficient to indicate if there were differences between the scoring formats related to expertise. In fact, the Knowledge and Experience variables had little if any predictive value. Alternate approaches to selecting test takers or to representing these aspects of expertise were considered and explored within the context of Study One.

A limitation to the Knowledge variable was the wide range in time between medical school graduation and taking the Part I examination that existed for test takers in these two studies. Canadian post-graduate trainees commonly completed the Part I examination in their final term of medical school while test takers from international post-graduate programs commonly attempt this examination as part of the process of applying for provincial licensure, which could be at anytime after graduation from medical school, a time span that ranges from one to 25 or more years. For test takers with a large gap between graduation and taking the Part I examination, Knowledge scores may have varied for reasons that would not correlate with their performance on individual patient cases. For example, loss of knowledge in areas outside their discipline and reduced recall of knowledge related to increasing age (Eva, 2002) are two factors that would negatively influence knowledge scores but would variably influence their clinical performance on specific cases (Hodges & McIlroy, 2003).

Two approaches to controlling for the unwanted variance in the Knowledge scores were briefly explored. In the first, the analyses for Study One were re-run using data from test takers where the gap between graduation and taking the Part I was less than 10 years. The result: Knowledge still did not contribute significantly to the predictions. Running a linear regression with the same model using total exam scores from the OSCE as the dependent variable also gave the same result. Coefficients for Knowledge were not

significant, meaning Knowledge scores from Part I were not predictive of performance-based scores. The second approach used to control for the variance in Knowledge scores related to the timing of the examination relative to medical school graduation was to select only Canadian post-graduate test takers for the analyses. The logistic regression studies were then re-run for Study One, after further selecting for test takers where there was rater agreement on the PIRS. None of the predictors were significant for the Depression instruments and only Skill score was significant for the Delirium instruments. Neither approach was conceptually satisfying as the study relied on having a cohort of test takers with disparate levels of clinical experience. Both approaches minimized this criterion and neither offered an improvement to the original model. The results were in keeping with the understanding that while formal learned knowledge is essential to expertise, recall of knowledge is not a predictor of expertise (Schmidt et al., 1990).

Experience was the weakest predictor and several of the analyses hinted that increasing years since graduation was associated with lower odds of passing, counter to what had been hypothesized when the variable was selected. Although time since graduation has been used in other studies to represent clinical experience (Dillon & Henzel, 1999; Sawhill, Dillon, Ripkey, Hawkins, & Swanson, 2003), these studies compared individuals who were still in training programs so that each year represented roughly the same amount of training for each test taker and the range of years since graduation was under ten years.

For the test takers in both studies, years since graduation had a less defined meaning, especially for international trainees. For test takers in these studies, years since graduation may have included years that were spent outside clinical training or practice and the range of years was large; running from less than a year to more than twenty-five years since graduation. For test takers in the upper range of years since graduation, the Experience variable, as with Knowledge, may have been capturing variance related to the negative effects of aging on clinical expertise (Eva, 2002) and for some, the negative impact of immigration and relocation challenges.

An alternate approach was considered to see if the Experience component of the model could be better represented. The type of problems each physician sees is largely related to their specialty and the process of specialization starts even before post-graduate training begins. Grouping test takers according to their specialty relative to each patient problem is therefore another way of representing relevant experience within an expertise model. The assumption is that test takers from specialties where the presenting patient problem is seen more frequently are more likely to have expertise relative to that clinical challenge and so Experience was replaced with Expert Category which was based on test taker specialty.

For Depression, family physicians are the most likely to be experienced with the patient problem. This patient might also be referred to psychiatry and while she is not a child, she is young and her current depression is related to the death of her father a year earlier and the impact this loss is having on her and the family. (She is living with her mother and younger siblings.) Given the family dynamic underlying the stress she is experiencing, pediatrics could also be considered a related specialty. So, for this case all family medicine, psychiatry and pediatric test takers were classified as being in a Related Specialty. Test takers from other specialties such as surgery, obstetrics or internal medicine were classified as being in an Unrelated Specialty. All other test takers, including those from laboratory specialties and pathology, were classified as Other and were assumed to be the least expert relative to the patient problem. Those whose training background was unknown were excluded from the analyses. For Depression, 76 test takers were removed from the data set.

For Delirium, where the patient is presenting with post-operative delirium secondary to alcohol withdrawal, the Related Specialties were deemed to be family medicine, internal medicine, surgery, and psychiatry. Test takers from disciplines like anaesthesia and obstetrics-gynecology were considered to be from Unrelated Specialties and test takers from pediatrics and specialties such as imaging, laboratory specialties and pathology were categorized as Other and were deemed the least expert group relative to

this patient problem. Forty-one test takers were excluded because their training background was unknown.

Four logistic analyses were run using Study One data. There was one analysis for each instrument for each case, using records where there was rater agreement on the PIRS. In brief, only Skill was a significant predictor across both cases. The Expert Category did not contribute to the model.

As a follow-up, two-factor ANOVA studies were run in SPSS 13.0 and the results for the four ANOVA studies are summarized in Table 32.

Table 32 *Two-factorial ANOVA for Study One*

Tests of Between-Subjects Effects for CL-1 Score (%)						
Source	Type III SS	*df*	Mean Square	*F*	Sig.	Partial η2
CPG-IPG	1247.131	1	1247.131	8.845	.003	.014
Expert Category	235.063	2	117.531	.834	.435	.003
CPG-IPG * Expert Category	618.706	2	309.353	2.194	.112	.007
Error	88404.008	627	140.995			
Total	3544225.000	633				
Tests of Between-Subjects Effects for CRS-1 Scores (%)						
Source	Type III SS	*df*	Mean Square	*F*	Sig.	Partial η2
CPG-IPG	4018.880	1	4018.880	19.920	.000	.031
Expert Category	734.547	2	367.273	1.820	.163	.006
CPG-IPG * Expert Category	421.443	2	210.722	1.044	.352	.003
Error	126499.228	627	201.753			
Total	3970300.000	633				
Tests of Between-Subjects Effects for CL-2 Score (%)						
Source	Type III SS	*df*	Mean Square	*F*	Sig.	Partial η2
CPG-IPG	.119	1	.119	.001	.980	.000
Expert Category	1239.517	2	619.759	3.416	.034	.012
CPG-IPG * Expert Category	181.632	2	90.816	.501	.606	.002
Error	98520.347	543	181.437			
Total	1990502.959	549				
Tests of Between-Subjects Effects for CRS-2 Score (%)						
Source	Type III SS	*df*	Mean Square	*F*	Sig.	Partial η2
CPG-IPG	3886.371	1	3886.371	14.571	.000	.026
Expert Category	1507.756	2	753.878	2.827	.060	.010
CPG-IPG * Expert Category	560.567	2	280.284	1.051	.350	.004
Error	144827.061	543	266.717			
Total	2852011.111	549				

The results showed a significant difference in scores between the Canadian and internationally trained test takers for three of the instruments (CL-1, CRS-1, CRS-2). The

reverse was true for CL-2 where the significant difference was between expert categories. An underlying problem was the imbalance between Canadian trainees and international trainees in the Expert Categories. Very few of the internationally trained test takers were in a Related Specialty, as indicated in Table 33.

Table 33 *Number of test takers by Expert Category for Study One*

	Depression		Delirium	
	Count			
Expert Category	**CPG**	**IPG**	**CPG**	**IPG**
Other	102	153	78	54
Unrelated Specialties	153	10	32	77
Related Specialties	209	6	297	11
Total	464	169	407	142

One-way ANOVA with test takers from Canadian programs showed no significant difference across categories for Skill or Knowledge, as seen in Table 34.

Table 34 *One-way ANOVA studies for Study One – CPG test takers*

One-way ANOVA for Study One – Depression						
		SS	***Df***	**Mean Square**	***F***	**Sig.**
Knowledge	Between Groups	22406.680	2	11203.340	1.908	.150
	Within Groups	2706880.369	461	5871.758		
	Total	2729287.050	463			
Skill Score (%)	Between Groups	139.234	2	69.617	2.667	.071
	Within Groups	12033.807	461	26.104		
	Total	12173.041	463			
Experience	Between Groups	126.142	2	63.071	6.690	.001
	Within Groups	4346.374	461	9.428		
	Total	4472.515	463			
One-way ANOVA for Study One – Delirium						
		SS	***Df***	**Mean Square**	***F***	**Sig.**
Knowledge	Between Groups	5509.548	2	2754.774	.473	.624
	Within Groups	2353926.477	404	5826.551		
	Total	2359436.025	406			
Skill Score%	Between Groups	64.441	2	32.221	1.382	.252
	Within Groups	9419.872	404	23.317		
	Total	9484.313	406			
Experience	Between Groups	65.823	2	32.912	3.436	.033
	Within Groups	3869.877	404	9.579		
	Total	3935.700	406			

There was a significant difference for Experience across the Expert Category groups for both cases with test takers from the Other category having a higher mean number of years of Experience, shown in Table 35.

Table 35 *Comparing Experience across Expert Categories for Canadian trainees*

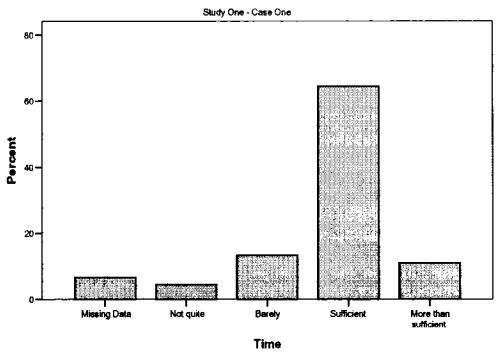
	Depression		Delirium	
	Experience[a]			
Expert Category	**N**	**Mean**	**N**	**Mean**
Other	102	2.76	78	2.77
Unrelated Specialties	153	1.34	32	1.72
Related Specialties	209	2.04	297	1.75
Total	464		407	

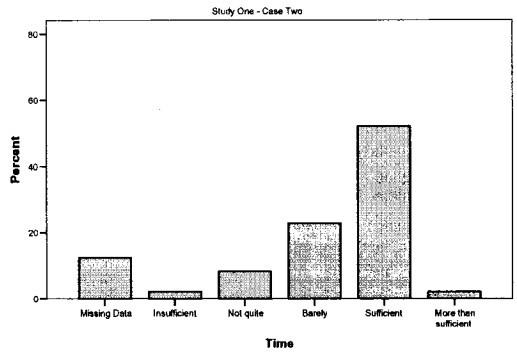
[a] **Experience is the number of years since graduation from medical school**

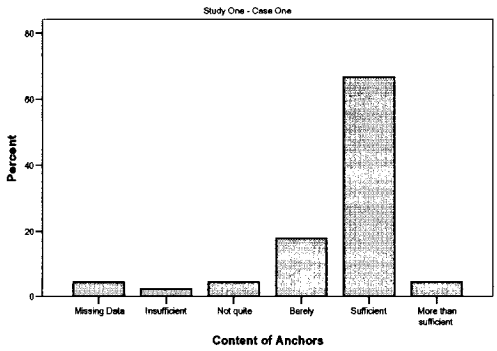
In sum, this means categorizing test takers by specialty was confounded by training background and years of experience. Those test takers with the most experience were least likely to have trained in a related specialty.

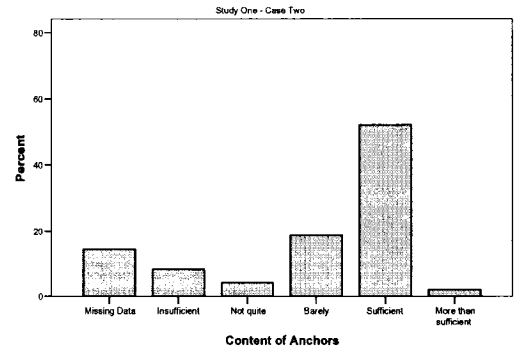
Experience was either a negative predictor of decisions or not a significant one in the regression analyses for both Study One and Two, contrary to what had been hypothesized when the variable was included. In a sense years since graduation were more representative of experience for these two studies than was immediately obvious since increasing years since graduation meant a test taker was less likely to have experience related to the specific patient problem with this cohort and more likely to be affected negatively by years out of clinical practice and age, in other words, less likely to be expert.

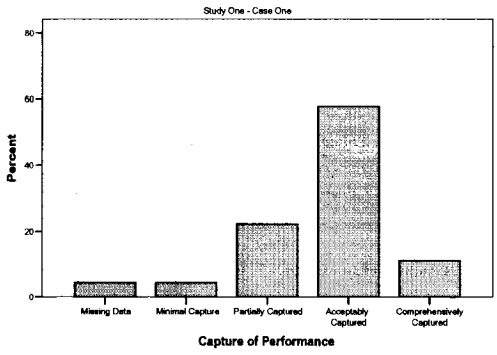
Appendix L: Study One - Bar Graphs from Assessor Responses

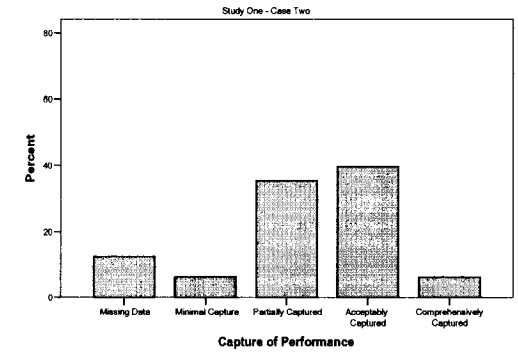


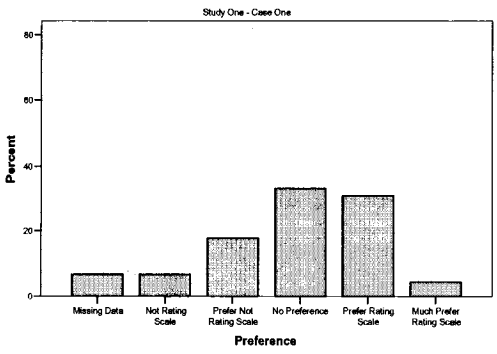


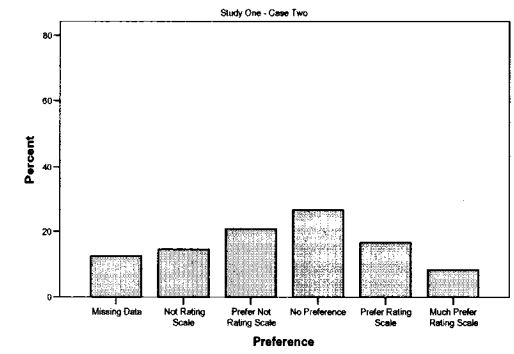






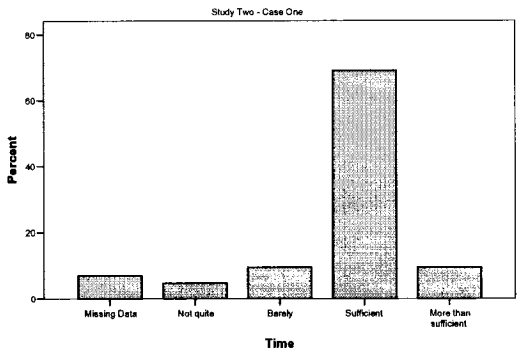


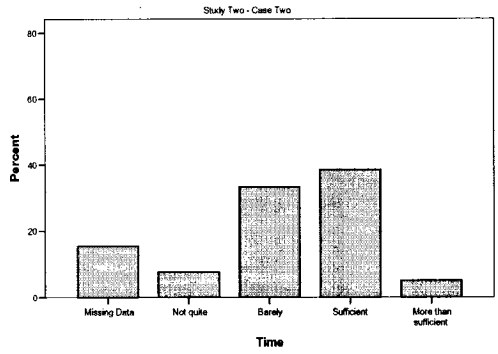


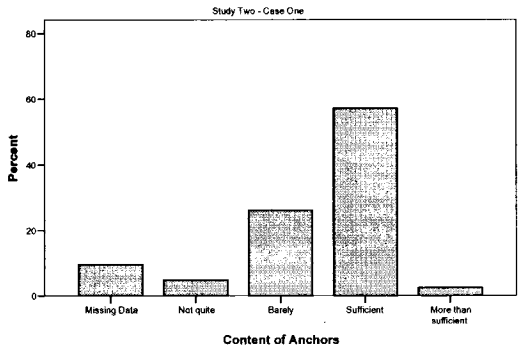


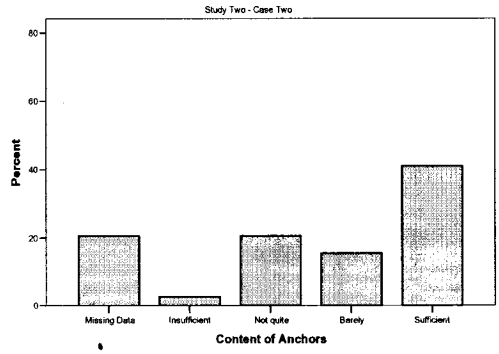
Note. Case One is Depression and Case Two is Delirium.

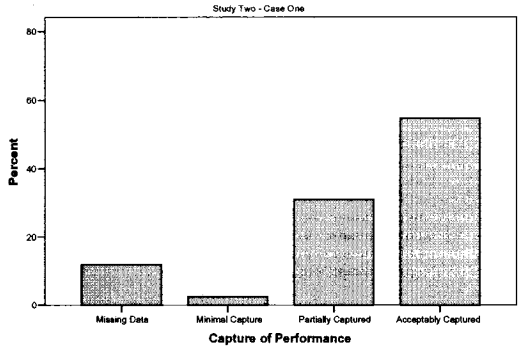
Appendix M: Study Two - Bar Graphs from Assessor Feedback

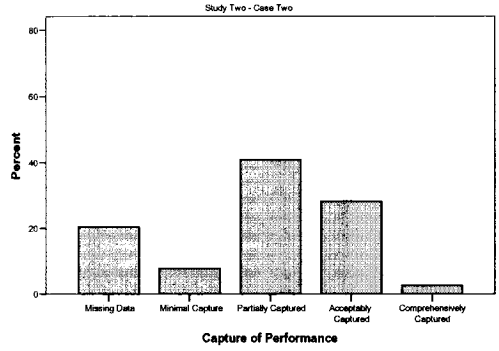


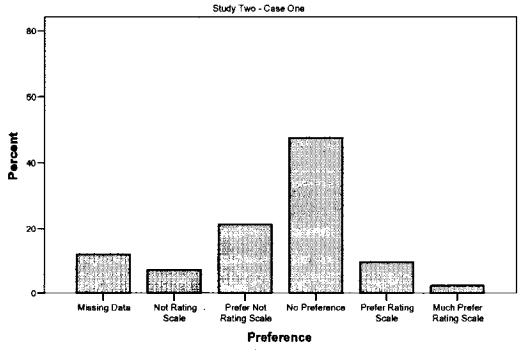


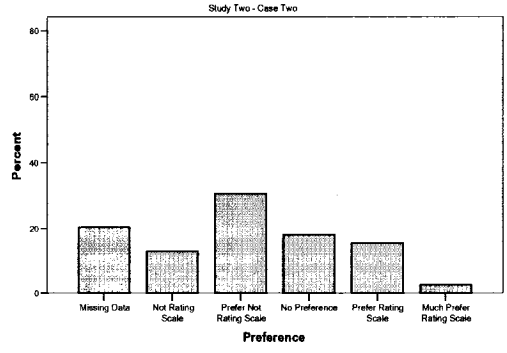












Note. Case One is Depression and Case Two is Delirium.