Adv in Health Sci Educ (2016) 21:609–625
DOI 10.1007/s10459-015-9655-4

In the minds of OSCE examiners: uncovering hidden assumptions

Saad Chahine[1] · Bruce Holmes[2] · Zbigniew Kowalewski[3]

Received: 8 April 2015 / Accepted: 23 November 2015 / Published online: 11 December 2015

Abstract The Objective Structured Clinical Exam (OSCE) is a widely used method of assessment in medical education. Rater cognition has become an important area of inquiry in the medical education assessment literature generally, and in the OSCE literature specifically, because of concerns about potential compromises of validity. In this study, a novel approach to mixed methods that combined Ordinal Logistic Hierarchical Linear Modeling and cognitive interviews was used to gain insights about what examiners were thinking during an OSCE. This study is based on data from the 2010 to 2014 administrations of the Clinician Assessment for Practice Program OSCE for International Medical Graduates (IMGs) in Nova Scotia. An IMG is a physician trained outside of Canada who was a licensed practitioner in a different country. The quantitative data were examined alongside four follow-up cognitive interviews of examiners conducted after the 2014 administration. The quantitative results show that competencies of (1) Investigation and Management and (2) Counseling were highly predictive of the Overall Global score. These competencies were also described in the cognitive interviews as the most salient parts of OSCE. Examiners also found Communication Skills and Professional Behavior to be relevant but the quantitative results revealed these to be less predictive of the Overall Global score. The interviews also reveal that there is a tacit sequence by which IMGs are expected to proceed in an OSCE, starting with more basic competencies such as History Taking and building up to Investigation Management and Counseling. The combined results confirm that a hidden pattern exists with respect to how examiners rate candidates. This study has potential implications for research into rater cognition, and the design and scoring of practice-ready OSCEs.

✉ Saad Chahine
saad.chahine@schulich.uwo.ca

[1] Department of Medicine and Faculty of Education, Centre for Education Research and Innovation, Schulich School of Medicine and Dentistry, University of Western Ontario, Health Sciences Addition, Room 114, London, ON N6A 5C1, Canada

[2] Division of Medical Education, Dalhousie University, Halifax, NS, Canada

[3] Faculty of Education, Mount Saint Vincent University, Halifax, NS, Canada

Keywords Assessment · Cognitive interview · Hierarchical Linear Modeling (HLM) · International Medical Graduate (IMG) · Objective Structured Clinical Exam (OSCE) · Practice-ready · Rater cognition

Introduction

The objective structured clinical exam (OSCE) is widely used in medical education assessments. Considerable research has been conducted to increase the reliability of such examinations. However, until recently few investigations have focused on associated validity issues. In this paper we report on a mixed methods approach that we used to probe more deeply into the complex processes underlying OSCE examinations.

Background

Since the late twentieth century, the use of OSCE examinations has become a major component of medical education assessment (Boursicot and Burdick 2014; Cox et al. 2007; Regehr et al. 2011). Statistical advancements such as those outlined by Brennan (2001), Cronbach et al. (1972), and Linacre and Wright (2002) have allowed researchers to estimate outcomes with more accuracy and streamline assessments (Khan et al. 2013; Medical Council of Canada 2013). Generally research in medical education assessment has been devoted to examining the reliability of assessments (Fuller et al. 2013; Hodges and McIlroy 2003; Hodges et al. 1999; Liao et al. 2010; Regehr et al. 1998). Recently researchers are interested more in the nature and validity of assumptions made by raters (Berendonk et al. 2012; Gingerich et al. 2014a, b; Johnston et al. 2013; Kogan et al. 2011; Wood 2014). Drawing on theories from social and cognitive psychology, emerging medical education research is focusing on understanding the underlying, hidden structures that govern rater cognition (Gingerich and Eva 2011). In parallel, researchers in the broader assessment community have moved a step further and developed conceptual models that can illustrate cognitive processes of raters when scoring open-ended items (Crisp 2012; Eckes 2012; Joe et al. 2011; Kishor 1990, 1995; Wolfe 2004, 2006). In this paper we describe a mixed methods approach to advance this effort in medical education. We anticipate the findings will provide insights particular to OSCE ratings and contributions to the larger assessment community regarding the tacit priorities of examiners rating performance for the workplace.

Hidden structures

Interpretation of results has always been an area of concern in assessment, due to the level of inference needed to build evidence towards the validity of conclusions about examinees (Cronbach and Meehl 1955; Kane 2013; Kane and Bejar 2014; Kelley 1927; Messick 1975; Shepard 1997). In performance assessment settings such as OSCEs, interpretations of IMGs results are made by examiners who are human. For instance, during a OSCE an examiner can quickly rate the History Taking skills of a candidate, provide a score, and move to the next candidate using a rating scale as a guide. However, if we were to

interview the examiner after the fact, some may provide general impressions that may or may not coincide with the ratings (Williams et al. 2003). We know little about the hidden structures that govern rater's cognition. However, in order to provide precise estimates of candidate performance, two elements at the very least are needed: congruent definitions of a construct (e.g. History Taking) and consistent measurement. The latter of these concerns can be easily addressed with the numerous psychometric models used to investigate reliability. The first relates to validity and requires in-depth investigations into structures, often latent, that govern rater decisions.

How raters decide what is important in an exam is invariably a hidden part of the assessment process. When ratings are produced, there is an assumption that the examiners' beliefs about what is important are universal and that these are accurately reflected in a candidate's score. However, researchers have found that identical examiner ratings may be based on different rationales (Douglas and Selinker 1992). Therefore, a major part of the justification of assessment procedures, not yet fully explored, is understanding what raters are thinking during the exam (Bejar 2012; Messick 1994). Knowledge of examiners' understanding of the assessment process is necessary to ensure the appropriateness of that process, the salience of the results and the utility of the OSCE in this context.

Methodology

Study setting

In North America, International Medical Graduates (IMGs) play a critical role in health care systems. Approximately one-quarter of family physicians have had some education outside North America (Boulet et al. 2009; Canadian Institute for Health Information 2009; Norcini et al. 2014; Walsh et al. 2011). In Canada, the assessment process for IMGs is provincially administered. The College of Physicians and Surgeons of Nova Scotia has one of the oldest IMG programs in Canada: the Clinician Assessment for Practice Program (CAPP). The CAPP OSCE differs from many other OSCEs, as it is a practice-ready assessment. A candidate has to pass a rigorous review process that includes the CAPP OSCE as well as a written therapeutics exam and file review. After the review process, successful IMGs receive a defined license requiring their practice to be mentored for 8 months and they must subsequently obtain certification from the College of Family Physicians of Canada within 4 years.

The examiners who participate in the CAPP OSCE are physicians currently practicing in Nova Scotia. Some are IMGs themselves and others have worked supervising IMGs. The CAPP OSCE comprises twelve, 12-min stations with 3-min intervals between the stations. All the stations are monitored and some are recorded during the exam for quality assurance. Over the past 8 years, the CAPP OSCE has evolved and developed to a high level of standardization and reliability from case development to implementation. The program has conducted several internal research studies to ensure consistency and optimize the exam (Maudsley 2008).

This study utilized data from CAPP OSCE administrations years 2010–2014. The quantitative data were examined along with four follow-up cognitive interviews of examiners after the 2014 administration. The interviews were intended to explore how examiners conceptualized practice-ready competence. The study was conducted with the

approval of the Mount Saint Vincent University Ethics Board and the College of Physicians and Surgeons of Nova Scotia.

Hierarchical Linear Modeling

Hierarchical Linear Modeling (HLM), is a multivariate statistical technique that is an extension of regression modeling developed in the early 1980s and designed to model nested data structures (Goldstein 1986; Raudenbush and Bryk 1986; Wong and Mason 1985). Since its development, the approach has become widespread and is utilized across multiple fields from economics to sociology and developmental psychology. Researchers have illustrated ways in which Logistic HLM can be used as a measurement model that is comparable to Rasch modeling (Kamata et al. 2008; Beretvas and Kamata 2005; Kamata 2001). The HLM approach lends itself to OSCE data as stations are nested within person and persons are nested within streams or cohorts. The HLM approach provides more accurate estimates of performance than traditional analysis methods because we do not have to aggregate or disaggregate data (Osborne 2000). In this study we used Ordinal Logistic Hierarchical Linear Modeling (OLHLM) to provide insight into which competencies are most predictive of Overall Global score. In doing so we are able to identify which competencies practicing family physicians believe are the most critical to being practice-ready.

Quantitative analyses

Data from 204 IMGs who participated in the CAPP OSCE in years 2010 though 2014 were used. In 2010 the OSCE had 14 stations. Since 2011 there have been 12 stations. Station cases are not repeated within a 3 year period. The CAPP did state that over the 5 years two cases were repeated; however, these were anonymized in the data set and as a result treated independently. At each station candidates were rated on a scale of 1 (Inferior), 2 (Poor), 3 (Borderline), 4 (Satisfactory), 5 (Very Good), and 6 (Excellent) on eight competencies:

(1) History Taking (HIST),
(2) Physical Exam (PE),
(3a) Physician Examiner Rated Communication Skills (PECOMM),
(3b) Simulated Patient Rated Communication Skills (SPCOMM),
(4a) Physician Examiner Rated Quality of Spoken English (PEQSE),
(4b) Simulated Patient Rated Quality of Spoken English (SPQSE),
(5) Counseling (COUN),
(6) Professional Behavior (BEHV),
(7) Problem Definition and Diagnosis (PDD), and
(8) Investigation and Management (INMAN).

They are also rated at each station on the outcome variable (Overall Global), which had the same 6-point scale. Since there were very few scores in the Inferior and Excellent categories, they were combined with adjacent scores for analysis purposes, resulting in four categories: Poor, Borderline, Satisfactory and Very Good. Due to ethical restrictions, the identity of the Examiners was randomly coded; as a result the researchers could not identify if an examiner was part of one or more years of administration. However, in consultation with the CAPP, they stated that some examiners return from year to year. In each year, all examiners participated in a mandatory orientation that included both online and face-to-face training components. SPSS 21 (IBM Corp., 2012), HLM 6.02

(Raudenbush et al. 2004) and the ggplot2 (Wickham and Chang 2015) package for R version 3.1.3 (CRAN 2015) were used to conduct the quantitative analyses.

Cognitive interviews

Asking participants to think aloud during a task is a common way to investigate the cognitive processes underlying tasks. However, due to the nature of an OSCE, we cannot interrupt examiners and ask them what they were thinking. Cognitive interviewing techniques were used in this study to elicit the thought processes of examiners (Willis 2005). These techniques are very flexible and may be applied in vivo or retrospectively. In the cognitive interview process, the interviewee conducts a task and is then asked to describe what they did and their thinking about why. The researcher in a cognitive interview can delve deeply and use probes to encourage the interviewee describe their thinking more comprehensively.

In the 2014 administration of the CAPP OSCE, Physician Examiners were asked to participate in an interview during which they watched a video of themselves in the exam room with the candidate and standardized patient. The video was paused every 2 min and each examiner was asked to describe their thinking. All of the participants ($n = 24$) in the exam were invited to participate in the study. However, due to ethics protocol requiring simulated patient, candidate and examiner consent, only 10 videos qualified to be used in the study. The examiners in the 10 videos were contacted and four agreed to take part in the interview. The remaining examiners did not respond to invitations. Those four received a $150 honorarium for their time.

The interviews were video recorded and transcribed verbatim. To conduct the analysis, the interview transcripts were divided into 12 two-minute segments. The segments were placed on a grid, to allow the researchers to look across the 12 segments and discern any patterns (Miles et al. 2014). We initially designed this study as mixed methods. The design was planned to be a parallel data collection with integration during interpretation, where the primary themes would be presented with comparable quantitative results (Creswell et al. 2011). However, due to the small sample, we were unable to draw an overall generalization or interpretation of the patterns from a thematic analysis. Instead, we explored specific excerpts that provide further insights into the quantitative results. Excerpts were chosen to be representative of similar interview stages across the participants.

Results

The quantitative data comprised of 2443 station level observations per competency and 204 (199 complete cases) candidates over 5 years. The Table 1 below provides the mean and standard deviation of each competency at the station level.

As shown in Table 1, the competencies were not assessed an equal number of times; for example, the Physical Exam (1214 station level observations) was present in approximately half of the stations, while History Taking was present in all. For every competency the standard deviation is approximately 1 point around the mean; thus, the majority of the scores were within one point of the mean.

The Overall Global outcome score (transformed to a 4-point scale) was not generated from an average of the individual competencies but was provided as a separate overall

Table 1 Descriptive statistics on competencies at each station

	Mean	Standard deviation	Station level: number of observations
HIST	4.14	0.97	2443
PE	4.01	0.96	1214
PECOMM	4.32	0.78	2421
SPCOMM	4.38	1.04	2426
PEQSE	4.41	0.75	2428
SPQSE	4.50	1.02	2429
COUN	3.93	1.01	1920
BEHV	4.52	0.75	2423
PDD	4.07	1.05	2359
INMAN	3.84	1.05	2425

These competencies are based on a 6-point Likert scale 1 (Inferior) to 6 (Excellent)

Data are from 199 complete cases over 5 years of administration

HIST History Taking, *PE* Physical Exam, *PECOMM* Physician Examiner Rated Communication Skills, *SPCOMM* Simulated Patient Rated Communication Skills, *PEQSE* Physician Examiner Rated Quality of Spoken English, *SPQSE* Simulated Patient Rated Quality of Spoken English, *COUN* Counseling, *BEHV* Professional Behavior, *PDD* Problem Definition and Diagnosis, *INMAN* Investigation and Management

judgment of the candidate's performance at that station. This variable does not appear in the table above as it was the outcome variable. There were 2415 complete observations of the Overall Global outcome that were used: 28 had missing values. Of the 2415 observations at the station level, 21 % were Poor, 33 % Borderline, 28 % Satisfactory, and 16 % Very Good. It is important to note the high-stakes nature of this exam and that candidates failed stations more often than they passed, which is commonplace in practice-ready IMG OSCEs (MacLellan et al. 2010).

Ordinal Logistic Hierarchical Linear Modeling results

The modeling procedure for OLHLM is similar to regression. The competencies were examined prior to inclusion in the model and were removed if not predictive, in order to achieve a parsimonious model. The variance–covariance matrix revealed that Problem Definition and Diagnosis (PDD), and Investigation and Management (INMAN) had a value close to 1, suggesting collinearity. We selected Investigation and Management (INMAN) to include in the model as it was a stronger predictor of the Overall Global score. Only four competencies [Professional Behavior (BEHV), Physician Examiner Rated Communication Skills (PECOMM), Counseling (COUN), and INMAN] were significant predictors. Other competencies: History Taking (HIST), Physical Exam (PE), and Quality of Spoken English (QSE) were not significant. The final model is presented in Appendix and Table 2 provides the results produced from the modeling.

The OLHLM results are based on the log-odds of four category outcomes using three connected equations, the first of which (see the Appendix) represents the lowest category (i.e. poor category). Since there was a smaller likelihood of a poor outcome than of the other categories combined, all the coefficients were negative. The greater the absolute magnitude of a competency coefficient, the more substantial it was as a predictor of

Table 2 Ordinal Logistic Hierarchical Linear Modeling results

	Regression coefficients (log-odds)	SE[a]	OR[b]	(95 % CI)[c]
Fixed effects				
Intercept (γ_{000})	−1.97*	0.58	0.14	(0.03, 0.70)
Station-level				
BEHV	−0.26**	0.08	0.77	(0.65, 0.90)
PECOMM	−0.28**	0.09	0.76	(0.64, 0.90)
COUN	−0.50***	0.07	0.61	(0.53, 0.69)
INMAN	−0.92***	0.06	0.40	(0.35, 0.45)
Threshold ($\delta_{(2)}$)	2.46***	0.09	11.75	(9.80, 14.11)
Threshold ($\delta_{(3)}$)	4.83***	0.13	125.12	(97.03, 161.35)
Random effects (Var. Variance components)				
Candidate level intercepts (r_{0jm})	0.20***			
Year level intercepts (u_{00m})	1.66***			

BEHV Professional Behavior, *PECOMM* Physician Examiner Rated Communication Skills, *COUN* Counseling, *INMAN* Investigation and Management

* $p < .05$; ** $p < .01$; *** $p < .001$

[a] SE: represents the standard error associated with each fixed effect

[b] OR: The odds ratios associated with each fixed effect. These values provide the relative amount by which the odds of belonging in the poor category decrease as the rating of a competency increases by 1

[c] 95 %CI: is the 95 % confidence interval of the odds ratios

Overall Global at the station level. The most predictive competency was Investigation and Management (INMAN, Coefficient = −0.92) and the least was Professional Behavior (BEHV Coefficient = −0.26).

The OLHLM is expanded upon in the Appendix. Using the results in Table 2 and functions 6–9 in the Appendix, we can derive probability estimates based on all candidates and stations. We can estimate that a given candidate had a 12 % chance of being in the poor category, 50 % chance of being borderline, 33 % chance of receiving a satisfactory and a 5 % chance of very good in a typical station. These estimates differ from the actual number of poor/borderline/satisfactory/very good categories, as they are based on the probability of an outcome from our parsimonious HLM model. The large coefficient on Investigation and Management indicates that in order to pass, a candidate had to score high on that competency.

Conceptualizing competence: introducing the interview data

To illustrate our results, we imputed plausible values (−3 to 3, at 0.1 intervals) in the model for each competency while holding the others constant (i.e. at zero). The following two figures illustrate the estimated probability of the Overall Global (vertical coordinate) by the score of an individual competence (horizontal coordinate) from low to high.

Figure 1 shows that the lines for each rating were fairly flat as ratings increased for Communication and for Professional Behavior. This suggests that it is difficult to predict

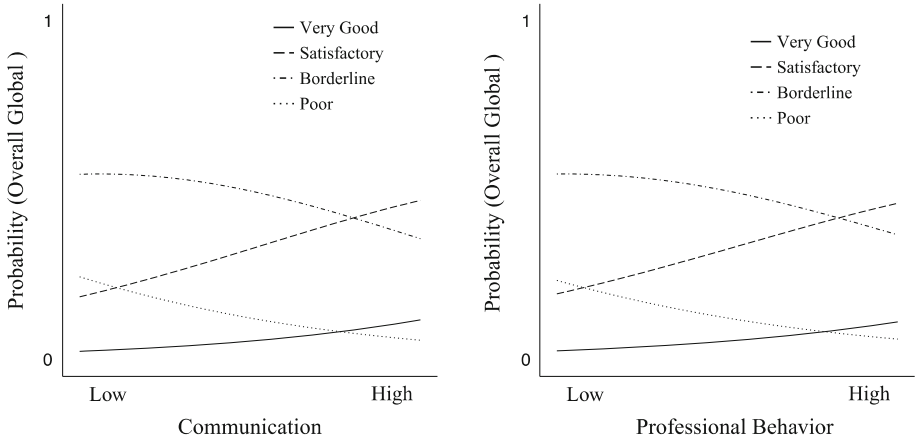

Fig. 1 Side by side graphs of estimated Overall Global outcome by ratings of Communication and Professional Behavior. *Note Each line* represents the change in probability of being in a given category (1. Very good, 2. Satisfactory, 3. Borderline, 4. Poor) as a function of the competency (i.e. Communication)

the likelihood of a candidate receiving any particular Overall Global outcome using just Communication or Professional Behavior. However we can see that as scores move from low to high, the satisfactory line ascends; these competencies appear to have been more predictive of satisfactory than of the other outcomes.

We can explore further why these two competencies are of relatively low value to examiners using the interview data. In his interview, Examiner 1 described both these competencies as interactions:

> Researcher: Can you describe, what do you mean by interaction?
> Examiner 1: You know the stuff that might want to build a rapport so that people can communicate to their patients, even getting to the meat of the substance. You know, the meet and greet, the sort of eye contact, how are you doing, what is your problem today.

Interactions that can be described as communication and professional behavior seem to be important at the very beginning of a 12-min OSCE. Examiner 4 described his/her thinking about a candidate who seems to be at the borderline level.

> Researcher: What are you thinking now?
> Examiner 4:
> … Probably the same what I was thinking at that time. Good questions. Some of them are certainly relevant for lab results and that's good. What's bad is, that we seem to be using the medical terms more than I would like for the patient of this nature, this type of a person. This supposed to be very straightforward, regular citizen who works at a diner and telling her "levels are high", she is just going to (the examiner is giving over the head gesture). You have to explain that. And it doesn't take long, you have to use simple words.

It was easy for examiners to discern when a candidate was not doing well when it came to communication. When communication was not an issue, an examiner began to evaluate other components of the OSCE. For example, Examiner 2 part way through the OSCE has already moved to History Taking:

Researcher: At this point can you describe what were you thinking?
Examiner 2:
So, in this case he took the history of what was going on with her and then I listened to see whether they'll branch out to do a comprehensive to cover the past medical history, current medications, any allergies, to cover a bit of social history, family history and a bit of a review of the system. So, he is progressing on those lines. So, right now I'm thinking; OK, he has finished off his history of what's going on with her, the history of the present illness and now he is moving on to past medical history, social history, medications etc. So, I'm thinking; all right this is reasonably well organized, his history taking. We can sort of see that they're going on the right track or not.

While all of the examiners described in detail good and poor qualities of the candidate's performance, they did so using a process: they were first looking for Communication and Professional Behavior, followed by History Taking. In the interview where they seemed to begin separating candidates into categories was about halfway through the OSCE, when they were expecting the candidates to start describing the issues at hand and to begin counseling the standardized patient. This is reflected in the quantitative data. The two figures below illustrate quantitative results on Counseling and Investigation and Management.

In comparison to Fig. 1 in which the lines were relatively flat, lines in Fig. 2 have much more movement as competency ratings shift from low to high. When it comes to Investigation Management, the very good category (solid line) ascends substantially as the scores increase and satisfactory begins to descend. Counseling, is also a strong predictor having more pronounced lines than those presented in Fig. 1.

During the OSCE and about halfway through (i.e. interview segments at 6 and 8 min), the examiners were looking for clues as to whether candidates have detected the problem through an investigative process and are managing the information gathered from the standardized patient. Examiner 2 reflected on her thinking during the station and described:

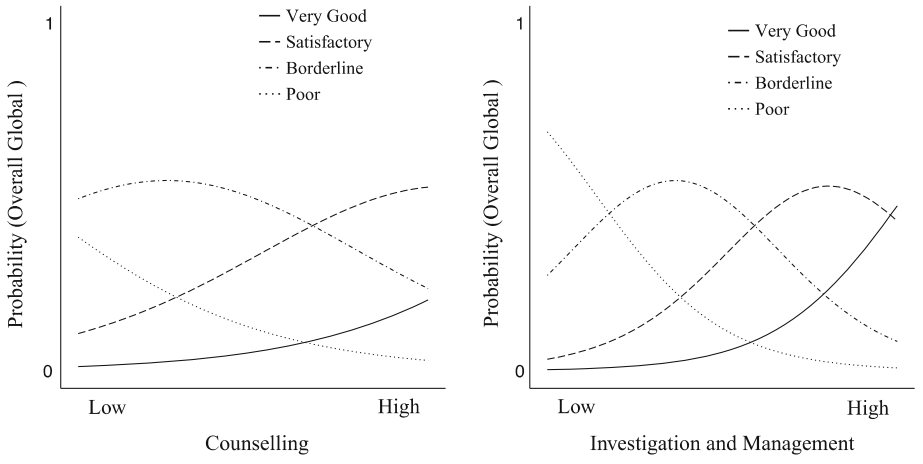

Fig. 2 Side by side graphs of estimated Overall Global outcome by ratings of Counseling and Investigation Management. *Note Each line* represents the change in probability of being in a given category (1. Very Good, 2. Satisfactory, 3. Borderline, 4. Poor) as a function of the competency (i.e. Counseling)

Researcher: What are you thinking now?
Examiner 2:
I remember thinking when he asked about her family history if she had any history of thyroid disease, I thought he's got it. So, that was good. He told her what he was thinking which I think is quite reasonable because I often tell patients what I think is going on before I examine them…So, clearly he's taking a good history, he is coming to a conclusion that she's got a thyroid problem.

Across the examiners, they were looking for candidates to narrow the possibilities to the most salient and important ones, making the clinical diagnosis and then translating that to a patient in a consultative, professional manner. When the candidate was doing well, it was qualitatively very different than when they were not doing as well. Examiner 4 describes this honing process as "selectivity" and described his thinking when the candidate did not perform this skill:

Researcher: What are you thinking now?
Examiner 4:
If he came here with an abdominal pain, you know, to me while you're listening to what they say, you examine the system that you think where the money is and then you start moving out. Not the other way around. Just throwing the net and hoping to catch something but that's not the way you work in the real world practice where you see patients every 5–10 min. You don't have 30 min per patient. So selectivity is a big thing. You have to be able to focus. This guy is not focused at all… Anyone can do that. And this is not what I would say is practice-ready.

In this quote from the last few minutes of the interview, the examiner was referring to a much more complex skill than investigation and management. All of the examiners suggested that physicians experience time pressures, especially in a rural setting. The examiner here was referring to the cognitive process that is expected of the physician: the ability to detect the problem, manage the information, and develop a treatment plan in a very short amount of time. Thus, there were two abilities examiners were looking for when assessing practice readiness: first being able to investigate and manage the information presented by the patient, and second to do this within the first 4–6 min of encounter.

Discussion

This paper combined an extended quantitative data set with focused, qualitative interviews in order to explore why examiners valued some things more than others in the CAPP OSCE. The quantitative evidence shows that examiners valued Investigation and Management above all other competencies. It requires the capture of information from a patient through an investigative process, the management of all the evidence, and formation of a conclusion. The examiners referred to it as "where the money is and then you start moving out" and "coming to a conclusion." In the minds of the examiners this complex skill of investigation and management is the "key feature" of the OSCE following the usage of Page et al. (1995). While there may have been individual differences when it comes to what was valued and when, and the often heterogeneous performance of IMGs, the results suggest that over 5 years of data, examiners are universally looking for the same competency manifestations in candidates. This echoes Gingerich et al.'s (2014b) recent

identification of a limited number of patterns of social judgments and their conclusion that examiners may not be as idiosyncratic as once theorized.

The qualitative evidence also suggests that examiners began to make and confirm their judgments halfway into the station. In addition to Investigation and Management, and Counseling skills, they were looking for the sequencing of the encounter to be paced in a such way that IMGs were able to make a diagnosis within 6–8 min and then to begin counseling a patient. Although more data is needed to confirm, there were hints to suggest that there is an expectation by examiners that practice-readiness amounts to the ability to diagnose accurately from a "thin slice" of information (Ambady and Rosenthal 1992).

This study expanded our current understanding of rater cognition by using a novel methodology; however, we wish to acknowledge several limitations. First, the quantitative data were collected from one province in Canada. Therefore, the findings may not generalize to other Canadian contexts. Further work is needed to understand whether similar findings can be replicated in other jurisdictions. Second, we were unable to account for examiners who repeatedly participated in the CAPP OSCE from 2010 to 2014. While we believe our overall model will be consistent, accounting for repeated examiners will provide for more accurate estimates of competency coefficients. And lastly the qualitative data were collected from four examiners. While we focused on themes that were consistently reported across all examiners, it is possible that with a larger sample size, we may have been able to capture additional themes or variability of beliefs and assumptions across the examiners based on their unique training and professional backgrounds.

Implications

What can we do with the knowledge that Investigation and Management (INMAN) was treated as the principal component by examiners of the CAPP OSCE? The conventional solution is to differentially weigh Investigation and Management to be worth more than other competencies. Weighting was developed almost a century ago by Toops (1927), and has since been applied across many fields (Bobko et al. 2007). However, there are two issues with the application of weights. The first is determining which competencies are worth more and which less in a composite score, and the second involves determining the magnitude of each weight. The first is not a simple task. Defining the ingredients of what practice-ready competency means for family medicine is challenging because the definition of *competent* and *competency* has been difficult to construct in a way that can be readily measured (Kane 1992; Williams et al. 2003; ten Cate et al. 2010; Epstein and Hundert 2002; Newble 2004; van der Vleuten 1996; van der Vleuten and Schuwirth 2005). The second task is equally complex from a mathematical point of view and a recent study in medical education reveals that establishing weights requires a very large data set and, when compared to non-weighted equivalents, weighting does not provide additional information (Sandilands et al. 2014). Considering the conceptual complexity of competency and the large data set required to derive weights, developing appropriate weights may not be feasible in OSCEs like the CAPP OSCE.

If the conventional solution of weighting is infeasible, we need to seek out other solutions. Another, simpler alternative for exams that are similar to the CAPP OSCE is to reorganize the rating scales by increasing complexity. Conventionally, the OSCE uses an anchored Likert rating scale where each competency is equally weighted. However, our results suggest an underlying hidden structure in the OSCE that lends itself to a Guttman style rating scale (Mislevy 1993). In a Guttman scale, tasks are ordered from easy to difficult, and success on the

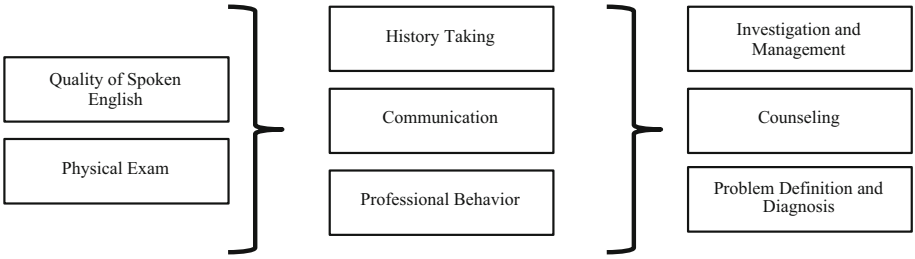

Fig. 3 Model of the nested structure of competencies for the CAPP OSCE

more difficult tasks implies success on the simpler ones. Visualized in Fig. 3 our results suggest that assessing competencies such as conducting a Physical Exam and Quality of Spoken English are more easy than competencies such as History Taking and Communication and Professional Behavior, which appear to be precursors to the more critical competencies of Problem Definition and Diagnosis, Investigation and Management and Counseling. This structure is hypothetical and based on the structures revealed in the data, which are intended to enhance the rating process for examiners. However, it is important to note that in practice physicians may be more fluid and natural in an encounter possibly doing some History Taking while performing a physical exam.

Although more research is needed to establish this structure, it is possible that the rating form could be reorganized so that the simpler competencies are assessed first using a simple rating scale, and the more complex are assessed using longer rating scales with areas for narrative. The reorganization to this scale may possibly reflect how examiners work, allowing them to direct their cognitive load where it matters—to the assessment of the principal competency of Investigation and Management.

Conclusion

The initial purpose of this study was to gain deeper insight into how examiners conceptualize practice-readiness. Through the combined use of OLHLM and cognitive interviews, we developed a novel way to capture what examiners believe are the essential abilities for medical practice. Quantitatively and qualitatively we found that the essence of "practice-ready" in the Nova Scotia context lay in the competencies of Investigation and Management along with some Counseling.

Despite the limitations of the study, we believe this work adds to the evidence base of rater cognition in medical education. This paper also suggests that future inquiries are needed to explore the generalizability of these findings, in other contexts, particularly with respect to how examiners prioritize competencies and their expectations relating to diagnostic efficiency. Lastly, this work suggests that in order to develop authentic assessment practices, we need to ensure that assessment measures and processes mirror the real world—that is, the ways in which examiners actually approach the assessment process and conceptualize competency.

Acknowledgments The authors wish to acknowledge the support of the College of Physicians & Surgeons of Nova Scotia in conducting the research study. The authors wish to acknowledge Susan Elgie, Dr. Lorelei Lingard, and Dr. Tomoko Arimura for their support in reviewing this manuscript.

Appendix

The Ordinal Logistic Hierarchical Linear Modeling (OLHLM) is slightly more complex as the outcome variable is an ordered categories and not continuous data. OLHLM is very similar to Logistic Hierarchical Linear Modeling, where the outcome is binary, and we estimating likelihood of receiving very good or poor ratings. Since there are four categories (poor, borderline, satisfactory, very good) of probability in our case three functions (1–3) are used and propensity of being in each of the categories is based on the change in log odds. The model below, represents a three level OLHLM with 4 category outcomes, 3 of the 4 categories are estimated, as it assumed that the 4th category is 1—the probability of being in the other three.

Station level model

$$\text{Category 1}: \quad \log\left[\frac{\phi'_{ijm(1)}}{1-\phi'_{ijm(1)}}\right] = \pi_{ojm} + \pi_{1jm}X_{1jm} + \pi_{2j}mX_{2jm} + \cdots + \pi_{kjm}X_{kjm} \tag{1}$$

$$\text{Category 2}: \quad \log\left[\frac{\phi'_{ijm(2)}}{1-\phi'_{ijm(2)}}\right] = \pi_{ojm} + \pi_{1jm}X_{1jm} + \pi_{2j}mX_{2jm} + \cdots + \pi_{kjm}X_{kjm} + \delta_{(2)} \tag{2}$$

$$\text{Category 3}: \quad \log\left[\frac{\phi'_{ijm(3)}}{1-\phi'_{ijm(3)}}\right] = \pi_{ojm} + \pi_{1jm}X_{1jm} + \pi_{2j}mX_{2jm} + \cdots + \pi_{kjm}X_{kjm} + \delta_{(3)} \tag{3}$$

Candidate level

$$\begin{aligned} \pi_{0jm} &= \beta_{00m} + r_{0jm} \\ \pi_{1jm} &= \beta_{10m} \\ \pi_{1jm} &= \beta_{10m} \\ &\vdots \\ \pi_{kjm} &= \beta_{k0m} \end{aligned} \tag{4}$$

Year level

$$\begin{aligned} \beta_{00m} &= \gamma_{000} + u_{00m} \\ \beta_{10m} &= \gamma_{100} \\ \beta_{20m} &= \gamma_{200} \\ &\vdots \\ \beta_{k0m} &= \gamma_{k00} \end{aligned} \tag{5}$$

Its important that when reading an HLM model to look at the different levels. In our case since there are no person characteristics at level 2 and no year characteristics at level 3, we are estimating the error variance at each of those levels and the competency coefficients. As such the:

$\phi'_{ijm(1)}$ is the probability that person j in year m scores a 1; $\phi'_{ijm(2)}$ is the probability that person j in year m scores a 1 or 2; $\phi'_{ijm(3)}$ is the probability that person j in year m scores a 1, 2 or 3; π_{ojm} is the intercept term; π_{kjm} is the coefficient for the kth competency for person j in year m; X_{kjm} is the kth competency score for person j in year m; $\delta_{(2)}$ is the threshold value between category 2 and 1; $\delta_{(3)}$ is the threshold value between category 3 and 2; r_{0jm} is the random component of π_{0jm}; β_{00m} is an effect of the reference competency in year m; β_{k0m} is an effect of the kth competency in year m; u_{00m} is the random component of β_{00m}; γ_{000} is the overall effects of competencies; γ_{k00} is the coefficient for competency k.

The HLM program uses the above model and provides estimates for each of the values above. However it does not provide the probability estimates of a candidate belonging to the poor, borderline, satisfactory or very good category at each station. These probability estimates need to be calculated. The following formulas are used to calculate the probability of being in each of the categories.

$$\text{Category 1}: \quad \phi'_{ijm(1)} = \frac{e^{(\gamma_{000}+\gamma_{100}X_{1jm}+\gamma_{200}X_{2jm}+\cdots+\gamma_{k00}X_{kjm})}}{1+e^{(\gamma_{000}+\gamma_{100}X_{1jm}+\gamma_{200}X_{2jm}+\cdots+\gamma_{k00}X_{kjm})}} \tag{6}$$

$$\text{Category 2}: \quad \phi'_{ijm(2)} = \left(\frac{e^{(\gamma_{000}+\gamma_{100}X_{1jm}+\gamma_{200}X_{2jm}+\cdots+\gamma_{k00}X_{kjm}+\delta_{(2)})}}{1+e^{(\gamma_{000}+\gamma_{100}X_{1jm}+\gamma_{200}X_{2jm}+\cdots+\gamma_{k00}X_{kjm}+\delta_{(2)})}}\right) - \phi'_{ijm(1)} \tag{7}$$

$$\text{Category 3}: \quad \phi'_{ijm(3)} = \left(\frac{e^{(\gamma_{000}+\gamma_{100}X_{1jm}+\gamma_{200}X_{2jm}+\cdots+\gamma_{k00}X_{kjm}+\delta_{(3)})}}{1+e^{(\gamma_{000}+\gamma_{100}X_{1jm}+\gamma_{200}X_{2jm}+\cdots+\gamma_{k00}X_{kjm}+\delta_{(3)})}}\right) - \phi'_{ijm(2)} - \phi'_{ijm(1)} \tag{8}$$

$$\text{Category 4}: \quad \phi'_{ijm(4)} = 1 - \phi'_{ijm(3)} - \phi'_{ijm(2)} - \phi'_{ijm(1)}\ (2) \tag{9}$$

$\phi'_{ijm(1)}$ denotes the probability of being in the poor category, when the X_{kjm} (i.e. competency scores) are at the average (i.e. set to zero). To calculate $\phi'_{ijm(4)}$ the probability of being in the very good category we subtract 1 from the probability of being in the satisfactory, borderline, or very good categories.

References

Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin, 111*(2), 256–274. doi:10.1037/0033-2909.111.2.256.

Bejar, I. I. (2012). Rater cognition: Implications for validity. *Educational Measurement: Issues and Practice, 31*(3), 2–9. doi:10.1111/j.1745-3992.2012.00238.x.

Berendonk, C., Stalmeijer, R. E., & Schuwirth, L. W. T. (2012). Expertise in performance assessment: assessors' perspectives. *Advances in Health Science Education, 18*, 559–571. doi:10.1007/s10459-012-9392-x.

Beretvas, S. N., & Kamata, A. (2005). The multilevel measurement model: Introduction to the special issue. *Journal of Applied Measurement, 6*(3), 247–254.

Bobko, P., Roth, P. L., & Buster, M. A. (2007). The usefulness of unit weights in creating composite scores: A literature review, application to content validity, and meta-analysis. *Organizational Research Methods, 10*(4), 689–709. doi:10.1177/1094428106294734.

Boulet, J. R., Cooper, R. A., Seeling, S. S., Norcini, J. J., & McKinley, D. W. (2009). U.S. citizens who obtain their medical degrees abroad: An overview, 1992–2006. *Health Affairs, 28*(1), 226–233. doi:10.1377/hlthaff.28.1.226.

Boursicot, K. A. M., & Burdick, W. P. (2014). Structured assessments of clinical competence. In T. Swanwick (Ed.), *Understanding medical education: Evidence, theory and practice* (2nd ed., pp. 293–304). New York: Wiley.

Brennan, R. L. (2001). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement, 38*(4), 295–317. doi:10.1111/j.1745-3984.2001.tb01129.x.

Canadian Institute for Health Information. (2009, August). *International Medical Graduates in Canada: 1972 to 2007 Executive Summary*. Retrieved February 1, 2015 from http://secure.cihi.ca/free_products/img_1972-2007_aib_e.pdf.

Corp, I. B. M. (2012). *IBM SPSS Statistics for Windows, Version 21.0*. Armonk, NY: IBM Corp.

Cox, M., Irby, D. M., & Epstein, R. M. (2007). Assessment in medical education. *New England Journal of Medicine, 356*(4), 387–396. doi:10.1056/NEJMra054784.

CRAN. (2015). *R 3.1.3 "Smooth Sidewalk"*. http://cran.r-project.org/.

Creswell, J. W., Klassen, A. C., Plano Clark, V. L., & Smith, K. C. (2011, August) for the Office of Behavioral and Social Sciences Research. Best practices for mixed methods research in the health sciences. National Institutes of Health. Retrieved August 1, 2015 from http://obssr.od.nih.gov/mixed_methods_research/pdf/Best_Practices_for_Mixed_Methods_Research.pdf.

Crisp, V. (2012). An investigation of rater cognition in the assessment of projects. *Educational Measurement: Issues and Practice, 31*(3), 10–20. doi:10.1111/j.1745-3992.2012.00239.x.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302.

Douglas, S., & Selinker, L. (1992). Analyzing oral proficiency test performance in general and specific purpose contexts. *System, 20*(3), 317–328. doi:10.1016/0346-251x(92)90043-3.

Eckes, T. (2012). Operational rater types in writing assessment: Linking rater cognition to rater behavior. *Language Assessment Quarterly, 9*, 270–292. doi:10.1080/15434303.2011.649381.

Epstein, R. M., & Hundert, E. M. (2002). Defining and assessing professional competence. *Jama, 287*(2), 226–235.

Fuller, R., Homer, M., & Pell, G. (2013). Longitudinal interrelationships of OSCE station level analyses, quality improvement and overall reliability. *Medical Teacher, 35*, 515–517. doi:10.3109/0142159X.2013.775415.

Gingerich, A., & Eva, K. W. (2011). Rater-based assessments as social judgments: Rethinking the etiology of rater errors. *Academic Medicine, 86*, S1–S7. doi:10.1097/ACM.0b013e31822a6cf8.

Gingerich, A., Kogan, J., Yeates, P., Govaerts, M., & Holmboe, E. (2014a). Seeing the "black box" differently: Assessor cognition from three research perspectives. *Medical Education, 48*, 1055–1068. doi:10.1111/medu.12546.

Gingerich, A., van der Vleuten, C. P. M., & Eva, K. W. (2014b). More consensus than idiosyncrasy: Categorizing social judgments to examine variability in Mini-CEX ratings. *Academic Medicine, 89*, 1510–1519. doi:10.1097/ACM.0000000000000486.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika, 73*(1), 43–56. doi:10.1093/biomet/73.1.43.

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

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

Joe, J. N., Harmes, J. C., & Hickerson, C. A. (2011). Using verbal reports to explore rater perceptual processes in scoring: A mixed methods application to oral communication assessment. *Assessment in Education: Principles, Policy & Practice, 18*, 239–258. doi:10.1080/0969594X.2011.577408.

Johnston, J. L., Lundy, G., McCullough, M., & Gormley, G. J. (2013). The view from over there: Reframing the OSCE through the experience of standardised patient raters. *Medical Education, 47*(9), 899–909. doi:10.1111/medu.12243.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement, 38*(1), 79–93. doi:10.1111/j.1745-3984.2001.tb01117.x.

Kamata, A., Bauer, D. J., & Miyazaki, Y. (2008). Multilevel measurement modeling. In A. A. O'Connell & D. B. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 345–390). Charlotte, NC: Information Age Publishing.

Kane, M. T. (1992). The assessment of professional competence. *Evaluation and the Health Professions, 15*(2), 163–182.
Kane, M. T. (2013). Validation as a pragmatic, scientific activity. *Journal of Educational Measurement, 50*(1), 115–122. doi:10.1111/jedm.12007.
Kane, M. T., & Bejar, I. I. (2014). Cognitive frameworks for assessment, teaching, and learning: A validity perspective. *Psicología Educativa, 20*(2), 117–123. doi:10.1016/j.pse.2014.11.006.
Kelley, T. L. (1927). *Interpretation of educational measurements*. New York: World Book Co. Retrieved February 1, 2014 from http://hdl.handle.net/2027/mdp.39015001994071.
Khan, K. Z., Gaunt, K., Ramachandran, S., & Pushkar, P. (2013). The objective structured clinical examination (OSCE): AMEE Guide No. 81. Part II: Organisation & Administration. *Medical Teacher, 35*(9), e1447–e1463. doi:10.3109/0142159X.2013.818635.
Kishor, N. (1990). The effect of cognitive complexity on halo in performance judgment. *Journal of Personnel Evaluation in Education, 3*, 377–386.
Kishor, N. (1995). The effect of implicit theories on raters' inference in performance judgment: Consequences for the validity of student ratings of instruction. *Research in Higher Education, 36*(2), 177–195. doi:10.1007/BF02207787.
Kogan, J. R., Conforti, L., Bernabeo, E., Iobst, W., & Holmboe, E. (2011). Opening the black box of clinical skills assessment via observation: A conceptual model. *Medical Education, 45*(10), 1048–1060. doi:10.1111/j.1365-2923.2011.04025.x.
Liao, S. C., Hunt, E. A., & Chen, W. (2010). Comparison between inter-rater reliability and inter-rater agreement in performance assessment. *Annals of the Academy of Medicine, Singapore, 39*(8), 613–618.
Linacre, J. M., & Wright, B. D. (2002). Construction of measures from many-facet data. *Journal of Applied Measurement, 3*(4), 486–512.
MacLellan, A.-M., Brailovsky, C., Rainsberry, P., Bowmer, I., & Desrochers, M. (2010). Examination outcomes for international medical graduates pursuing or completing family medicine residency training in Quebec. *Canadian Family Physician, 56*(9), 912–918.
Maudsley, R. (2008). Assessment of international medical graduates and their integration into family practice: The clinical assessment for practice program. *Academic Medicine, 83*, 309–315.
Medical Council of Canada. (2013, November). *Guidelines for the development of objective structured clinical examination (OSCE) cases*. Retrieved February 1, 2015, from http://mcc.ca/wp-content/uploads/osce-booklet-2014.pdf.
Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist, 30*(10), 955–966. doi:10.1037/0003-066X.30.10.955.
Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13–23. doi:10.3102/0013189X023002013.
Miles, M. B., Huberman, A. M., & Saldana, J. (2014). *Qualitative data analysis: A methods sourcebook* (3rd ed.). Thousand Oaks: Sage.
Mislevy, R. J. (1993). Foundations of a new test theory. In N. Frederikson, R. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 19–49). Hilllsdale, NJ: Lawrence Erlbaum Associates.
Newble, D. (2004). Techniques for measuring clinical competence: Objective structured clinical examinations. *Medical Education, 38*(2), 199–203. doi:10.1046/j.1365-2923.2004.01755.x.
Norcini, J. J., Boulet, J. R., Opalek, A., & Dauphinee, W. D. (2014). The relationship between licensing examination performance and the outcomes of care by international medical school graduates. *Academic Medicine, 89*, 1157–1162. doi:10.1097/ACM.0000000000000310.
Osborne, J. W. (2000). Advantages of hierarchical linear modeling. *Practical Assessment, Research & Evaluation, 7*(1). Retrieved February 6, 2015 from http://PAREonline.net/getvn.asp?v=7&n=1.
Page, G., Bordage, G., & Allen, T. (1995). Developing key-feature problems and examinations to assess clinical decision-making skills. *Academic Medicine, 70*(3), 194.
Raudenbush, S., & Bryk, A. S. (1986). A hierarchical model for studying school effects. *Sociology of Education, 59*(1), 1–17. doi:10.2307/2112482.
Raudenbush, S. W., Bryk, A. S., & Congdon, R. (2004). *HLM 6 for Windows*. Skokie, IL: Scientific Software International, Inc.
Regehr, G., Eva, K., Ginsburg, S., Halwani, Y., & Sidhu, R. (2011). *Assessment in postgraduate medical education: Trends and issues in assessment in the workplace* (Members of the FMEC PG consortium). Retrieved February 1, 2015 from https://www.afmc.ca/pdf/fmec/13_Regehr_Assessment.pdf.
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*(9), 993–997.

Sandilands, D. D., Gotzmann, A., Roy, M., Zumbo, B. D., & de Champlain, A. (2014). Weighting checklist items and station components on a large-scale OSCE: Is it worth the effort? *Medical Teacher, 36*(7), 585–590. doi:10.3109/0142159X.2014.899687.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16*(2), 5–24. doi:10.1111/j.1745-3992.1997.tb00585.x.

ten Cate, O., Snell, L., & Carraccio, C. (2010). Medical competence: The interplay between individual ability and the health care environment. *Medical Teacher, 32*(8), 669–675. doi:10.3109/0142159X.2010.500897.

Toops, H. A. (1927). The selection of graduate assistants. *Personnel Journal (Pre-1986), 6*, 457–472.

van der Vleuten, C. P. M. (1996). The assessment of professional competence: Developments, research and practical implications. *Advances in Health Science Education, 1*(1), 41–67. doi:10.1007/BF00596229.

van der Vleuten, C. P. M., & Schuwirth, L. W. T. (2005). Assessing professional competence: From methods to programmes. *Medical Education, 39*(3), 309–317. doi:10.1111/j.1365-2929.2005.02094.x.

Walsh, A., Banner, S., Schabort, I., Armson, H., Bowmer, M. I., & Granata, B. (2011). *International Medical Graduates—Current issues* (Members of the FMEC PG consortium). Retrieved February 1, 2015 from http://www.afmc.ca/pdf/fmec/05_Walsh_IMG%20Current%20Issues.pdf.

Wickham, H., & Chang, W. (2015). *Ggplot2: An implementation of the grammar of graphics, Version 1.0.1*. http://cran.r-project.org/web/packages/ggplot2/index.html.

Williams, R. G., Klamen, D. A., & McGaghie, W. C. (2003). Cognitive, social and environmental sources of bias in clinical performance ratings. *Teaching and Learning in Medicine, 15*(4), 270–292. doi:10.1207/S15328015TLM1504_11.

Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks: Sage.

Wolfe, E. W. (2004). Identifying rater effects using latent trait models. *Psychology Science, 46*(1), 35–51.

Wolfe, E. W. (2006). Uncovering rater's cognitive processing and focus using think-aloud protocols. *Journal of Writing Assessment., 2*(1), 37–56. http://www.journalofwritingassessment.org/archives/2-1.4.pdf.

Wong, G. Y., & Mason, W. M. (1985). The hierarchical logistic regression model for multilevel analysis. *Journal of the American Statistical Association, 80*(391), 513–524. doi:10.2307/2288464.

Wood, T. J. (2014). Exploring the role of first impressions in rater-based assessments. *Advances in Health Science Education, 19*, 409–427. doi:10.1007/s10459-013-9453-9.