Running head

Running head: Validity argument for a simulated assessment

Title

Constructing and Evaluating a Validity Argument for the Final-Year Ward Simulation Exercise

Hettie Till, MSc, MMedEd, DEd.

Jean Ker, MD, FRCGP, FRCPE, FHEA

Carol Myford, MA, PhD.

Kevin Stirling, MSc, FHEA

Gary Mires, MD, FRCOG, FHEA

Dr Till is a consultant in assessment and holds affiliate membership in the Centre for Medical Education, School of Medicine, University of Dundee, UK.

Dr Ker is Associate Dean of Innovation in Medical Education and Professor of Medical Education at the School of Medicine, University of Dundee, UK

Dr Myford is an associate professor in the Department of Educational Psychology, College of Education, University of Illinois at Chicago, USA.

Mr Stirling is a lecturer in simulation at the Clinical Skills Centre, School of Medicine, University of Dundee, UK.

Dr Mires is Dean of Medicine and Professor of Obstetrics, School of Medicine, University of Dundee, UK.

Corresponding author: Hettie Till

11 Van Riebeeck Street

Franschhoek

South Africa

7690

Tel: +27 (0)21 876 4035

e-mail: hettietill@gmail.com

Key words

Validity argument

Validity propositions

Validity evidence

Score interpretation

Clinical performance

Clinical ability

Simulated ward

Multi-facet Rasch measurement

Generalizability theory

Simulation in assessment of clinical performance

Evaluating medical students' fitness to practice

Abstract

The authors report final-year ward simulation (FYWSE) data from the University of Dundee Medical School. Faculty who designed this assessment intend for the final score to represent an individual senior medical student's level of clinical performance. The results are included in each student's portfolio as one source of evidence of the student's capability as a practitioner, professional, and scholar. Our purpose in conducting this study was to illustrate how assessment designers who are creating assessments to evaluate clinical performance might develop propositions and then collect and examine various sources of evidence to construct and evaluate a validity argument.

The data were from all 154 medical students who were in their final year of study at the University of Dundee Medical School in the 2010 – 2011 academic year. To the best of our knowledge, this is the first report on an analysis of senior medical students' clinical performance while they were taking responsibility for the management of a simulated ward.

Using multi-facet Rasch measurement and a generalizability theory approach, we examined various sources of validity evidence that the medical school faculty have gathered for a set of six propositions needed to support their use of scores as measures of students' clinical ability. Based on our analysis of the evidence, we would conclude that, by and large, the propositions appear to be sound, and the evidence seems to support their proposed score interpretation. Given the body of evidence collected thus far, their intended interpretation seems defensible.

Introduction

The valid and reliable assessment of clinical performance at any stage of medical practice is important for the reassurance of patients and for the accountability of the profession (General Medical Council, 2010). In the United Kingdom (UK) towards the end of undergraduate medical education, faculty must determine whether medical students have acquired the necessary level of clinical ability to enable them to provide high quality care for their patients in the context of the National Health Service (NHS). Providing sound validity evidence to support their claim that students have demonstrated this level of clinical ability can be challenging (Epstein, 2007).

Building a convincing validity argument to support the interpretations that assessment designers and assessment users make from the scores on a performance assessment is a formidable undertaking (Kane, 2006, 2013; Messick, 1989, 1996; Miller & Linn, 2000). According to the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014), the validation process begins with an explicit statement of how assessment designers intend for scores on that assessment to be interpreted and used (Standard 1.1), along with a rationale for each intended interpretation. Providing a rationale involves identifying the set of propositions (or claims) needed to support an intended score interpretation for a particular assessment use (Standard 1.2).

The assessment designers must then develop validity arguments that support (or refute) each intended interpretation of scores for a specified use. That is, they must determine which type(s) of validity evidence (e.g., evidence based on test content, response processes, internal structure, relations to other variables, and/or consequences of testing) are most needed for evaluating the soundness of each proposition, and then collect and examine that evidence to determine to what extent it supports the proposed score interpretation.

As the *Standards* note, validation is an ongoing process with additional evidence gathered and integrated into the validity argument as it becomes available. However, at various points in time, assessment designers must present summaries of the available evidence so that assessment users can determine whether the evidence is sufficient to support the intended score interpretations in their particular settings (Standard 1.2).

What might this validation process look like for an assessment of medical students' clinical performance if the assessment designers were to use the newly revised *Standards for Educational and Psychological Testing* to guide them? The assessment designers would first establish how they intended scores on that assessment to be interpreted (e.g., as a measure of an individual student's level of clinical ability) and used (e.g., to identify students who have acquired the level of clinical ability needed to provide high quality patient care). They would define the clinical ability construct they are measuring and describe the particular aspects of clinical ability that scores on the assessment are intended to represent, carefully distinguishing clinical ability from other related constructs such as prioritization, response to interruptions or communication. They would then provide a rationale for their intended score interpretation by stating a set of propositions that would support using the medical students' scores for the purpose they had identified.

When they are developing their set of propositions, the assessment designers would need to identify various facets of that clinical performance assessment context that might threaten the validity of the score interpretation. For example, if multiple assessors were involved, then the assessment designers would likely include propositions about those assessors and the quality of their judgments (e.g., the assessors have appropriate backgrounds and experience to carry out the clinical performance assessment, the assessors are adequately trained, the assessors score accurately and reliably, the assessors' scoring standards do not drift over time). Additionally, if the assessment were composed of a sample of exercises from a content domain, there would also likely be some propositions about those exercises (e.g., the sample of exercises adequately represents the content domain of the assessment, performance on the sample of exercises can be generalized to the content domain that is being assessed), as well as propositions about the students who engaged in those exercises (e.g., the students responded to the exercises in the manner that the assessment designers intended, the students were fairly assessed, the students who demonstrate competent performance on the assessment will be able to provide high quality patient care).

Once the assessment designers had identified their set of propositions, they would then collect and examine various sources of validity evidence to evaluate the soundness of each proposition. That is, they would build a validity argument, integrating strands of evidence to determine the extent to which the evidence they have collected supports using students' scores to identify those who have acquired the level of clinical ability needed to provide high quality patient care.

Purpose of the Study

In this study, we illustrate how assessment designers who are creating assessments to evaluate medical students' clinical performance might approach the tasks of developing propositions and collecting and examining various sources of validity evidence. We use as our example the final-year Ward Simulation Exercise (FYWSE), created by faculty and the University of Dundee Medical School to assist them in determining whether their senior medical students have acquired the level of clinical ability needed to provide high quality patient care.   The results obtained from the administration of the FYWSE are included in each student's portfolio as one source of evidence of the student's capability as a practitioner, professional, and scholar. These results, along with other sources of evidence gathered at various points during the student's training program, function as the final portfolio examination prior to graduation (General Medical Council, 2009).

We posed a series of research questions about several key facets of the FYWSE assessment that have the potential to introduce unwanted construct-irrelevant sources of error into students' scores: the assessors who evaluated the students' clinical performance, the exercises included in this assessment, the domains that were measured, and the medical students themselves. Answers to these research questions provided sources of validity evidence related to a set of propositions that are central to the validity argument for this assessment. In our investigation, we used several statistical approaches to obtain empirical evidence for the propositions. We present the evidence for each proposition and then determine to what extent the evidence supports the proposed score interpretation. We conclude by considering the implications of our research for the ongoing FYWSE validation process.

We chose to use two complementary approaches to analyze the rating data from this administration of the FYWSE: a generalizability theory (g-theory) approach, and a multi-facet Rasch measurement (MFRM) approach. While the two approaches share some common characteristics, they also differ in some important ways. Other researchers have documented key differences between these two statistical approaches in how they conceptualize and handle measurement error (e.g., Eckes, 2011; Iramaneerat, Yudkowsky, Myford, & Downing, 2008; Linacre, 1996, 2001; Lynch & McNamara, 1998; Smith & Kulikowich, 2004; Sudweeks, Reeve, & Bradshaw, 2005). In this next section, we explain how assessment designers can use these approaches to obtain various sources of validity evidence. Both approaches will provide evidence that assessment designers can employ to help build validity arguments, but they will use the evidence to evaluate the soundness of different sets of propositions.

Assessment designers could use the results from a g-theory analysis to investigate propositions that focus on the generalizability or dependability of scores. When creating the rationale to support their intended interpretation of scores on a clinical performance assessment, they might pose propositions to argue that scores can be generalized across alternate forms of the assessment, different administrations of the assessment, different subgroups of students taking the assessment, and different assessors who rate students' performances. They might also pose propositions to argue that the number of items/domains included in the assessment and the number of assessors who rate students' performances are sufficient to produce dependable, reliable/precise scores across replications of the assessment procedure.

The assessment designers could then use a g-theory approach to analyze the rating data from their clinical performance assessment, and gather evidence to allow them to evaluate the soundness of their propositions. From the results of their preliminary G study, they could examine validity evidence for propositions related to score generalizability, determining the impact of each of these sources of variability on scores. Using analysis of variance (ANOVA) procedures, they could obtain estimates of the variance components associated with each error source and then compare them to identify the major sources of error (i.e., determine what the contribution of each source of error is to the overall measurement error). From the results of their subsequent D study, they could examine validity evidence for propositions related to score dependability. If they found that the evidence from this study suggested that one (or more) of the propositions were not sound, they could then consider various strategies that might prove effective in decreasing the variance in scores due to measurement error, thereby increasing the reliability/precision of those estimates (i.e., compare and contrast the effects of increasing the number of scoring opportunities included in the assessment (as opposed to increasing the number of assessors) on the reliability/precision of the resulting scores).

While a g-theory analysis provides useful information about group-level main effects of various facets in an assessment procedure and the interactions of those facets, a MFRM analysis checks on the functioning of individual elements within each facet of that procedure (e.g., examining the performance of individual exercises, items/domains, students, assessors, rating scales, etc.). When creating the rationale to support their intended interpretation of scores on the clinical performance assessment, assessment designers could use the results from a MFRM analysis to investigate a qualitatively different set of propositions. For example, they might pose propositions about the performance of the assessors (e.g., The assessors exercise similar levels of severity when rating students' performance on the simulation exercises. Any differences in assessor severity do not affect students' scores.). They might also pose propositions about the assessment instrument (e.g., Each 5-point rating scale included on the instrument functions as a 5-point scale. The items/domains work together to measure clinical ability.), and about the students (e.g., The assessment procedure is effective in identifying those students who have acquired the necessary level of clinical performance to provide high quality patient care, and those who still need more practice).

If the assessment designers use a MFRM approach to analyze the rating data from their clinical performance assessment, they could gather evidence to allow them to evaluate the soundness of their propositions. The MFRM approach logistically transforms assessors' ordinal ratings to an equal-interval logit scale of measures. Technically, the dependent variable is the logistic transformation of ratios of successive rating scale category probabilities (log odds), and the independent variables are the different facets of the assessment procedure that affect measurement. In practice, this means that when a MFRM analysis is run, the various facets are analyzed simultaneously but independently and calibrated onto a single linear scale. The results from a MFRM analysis provide much needed diagnostic information about the impact of different sources of error on the quality of assessors' ratings. If the MFRM results suggested that one (or more) of the propositions were not sound, then the assessment designers could use their results to pinpoint the individual elements of their assessment procedure that were not functioning as intended (i.e., establish quality control). After they identify those problematic elements, they would be in a good position to consider possible next steps they might take to improve the fairness and objectivity of their procedure (Myford & Wolfe, 2003).

Background

For a number of years, medical schools have focused increasing attention on the use of workplace-based assessment (WPBA) tools to examine clinical performance (Norcini, 2003, 2005; Norcini & McKinley, 2007; Postgraduate Medical Education and Training Board and Academy of Medical Royal Colleges, 2009). The use of WPBA has also increased in the UK at both undergraduate and postgraduate levels (Hayes, 2011; Mcleod, Mires, & Ker, 2012). However, the validity and reliability of scores from different forms of WPBA are contested. McKinley et al. (2008) argued "that existing checklists frequently do not enable assessment of humanistic and teamwork competencies" (p. 42) and scores may not reflect capabilities (Bindal, Wall, & Goodyear, 2011). Morris, Gallagher, and Ridgway (2012) noted low inter-rater reliability in a number of the eight published tools for assessing medical students' use of procedural skills. Medical educators have questioned both the feasibility of using some of the WPBA tools for high-stakes assessment (Murphy, Bruce, Mercer, & Eva, 2009) and the variable contexts used for providing feedback to students (Miller & Archer, 2010).

The use of simulation (Issenberg & Scalese, 2007; Ker & Bradley, 2014; Maran & Glavin, 2007) as an alternative to WPBAs to safely assess clinical performance within an authentic setting (Schuwirth & van der Vleuten, 2003) has potentially many benefits. These include improving patient safety, patient handover, measuring task allocation (Siassakos et al., 2011), and enhancing multiprofessional team working (Crofts et al., 2007). Studies have measured the benefit of simulated learning to workplace performance in the context of surgery and in catheter-acquired infections (Barsuk, Cohen, McGaghie, & Wayne, 2010; Grantcharov et al., 2004; Sturm et al., 2008). As part of the assessment process, simulation can also provide a safe opportunity for students to receive feedback on their performances (Stiggins, 2005; Stiggins & Chappuis, 2006). One of the challenges in simulation-based assessment, however, is that there is currently little validity evidence to support its use (Cook, Brydges, Zendejas, Hamstra, & Hatala, 2013).

Following a feasibility study in 2008, the University of Dundee Medical School introduced a final-year ward simulation exercise (FYWSE) in 2009 to objectively assess the clinical performance of senior medical students in their last year of training, providing validity evidence of graduating students' readiness for practice (General Medical Council, 2011). The FYWSE is an integral part of the undergraduate medical curriculum. Faculty designed an assessment tool to rate the medical students' clinical ability as they cared for six patients on a simulated ward for a 20-minute period. Faculty based the FYWSE design on a utility index framework (van der Vleuten & Dannefer, 2012; van der Vleuten & Schuwirth, 2005). They were drawn to this framework out of a desire to ensure that their assessment would provide consistent results (reliability) that were trustworthy (validity). They wanted to be able to provide evidence of performance (educational impact) without compromising patient care (practicability).

Research Questions and Associated Propositions for Building a Validity Argument

In this paper, we share our exploration of the FYWSE, an assessment instrument that the University of Dundee Medical School is using to identify students who have acquired the level of clinical ability needed to provide high quality patient care. Faculty who designed this assessment intend for the final score to represent an individual senior medical student's level of clinical performance.

Listed in Table 1 are a set of propositions that the assessment designers could use to begin to build a validity argument to support the use of scores on the FYWSE for its intended purpose. To the right of each proposition are the research questions that we posed to provide validity evidence related to each proposition.

Method

Context

Development of the FYWSE. When designing the simulation assessment, the medical faculty invited a group of junior doctor supervisors to participate in a focus group to identify clinical performance behaviours that foundation (junior) doctors would need to demonstrate in order to make safe decisions, communicate effectively, perform clinical tasks and prioritize their workload. The medical faculty mapped these identified behaviours against the Tomorrows Doctors (General Medical Council, 2011) and Foundation Programme curriculum outcomes (Foundation Programme 2012) and, where appropriate, linked each behaviour to best evidence medicine and use of protocols and guidelines (Ker et al., 2009; McIlwaine, McAleer, & Ker, 2007).

The FYWSE was developed to replicate and reflect the reality of the clinical workplace. Medical faculty used an iterative process (Linstone & Turoff, 1975) to design the assessment, convening focus groups of health care practitioners who work with junior doctors, reviewing admissions to an acute medical unit and commonly reported errors, observationally shadowing junior doctors to identify common interruptions and task management issues, and conducting a literature review of common presentations. This process, which involved medical education and clinical experts, identified three main elements related to patient care reflective of junior doctors' workload and the outcomes of the Foundation curriculum and Tomorrows Doctors (General Medical Council, 2011): a new admission, communication issues, and the management of the acutely unwell patient. Faculty identified the new admissions from a review of the acute admissions to the medical unit in the hospital and based their selection of new admissions on the frequency of their presentations. To create an authentic context for assessment that reflects the complex, real-world pressures that physicians face in the clinical workplace, trained simulated patients participate in the ward simulation exercises, and the trainers used to teach them the required invasive procedures are very realistic. Additionally, as students are providing patient care on the simulated ward, they are intentionally interrupted by patients and health care practitioners - through the pager system, personally, or face to face (Ker, Hesketh, Anderson, & Johnson, 2005, 2006). It is recognized through our increased understanding of complexity theory that the workplace can become unsafe for practitioners when they are under pressure. The ward simulation exercises were thus developed to reflect the pressures that occur in the clinical workplace.

Using a modified Delphi technique, medical faculty identified the domains to be included in the FYWSE, basing their decisions on the GMC Tomorrows Doctors Good Medical Practice (General Medical Council, 2009). Initially, six domains were identified, developed and piloted, but the assessors found that they needed more than six domains to adequately rate the distinctly different tasks upon which the students were judged. A subsequent iterative process involving clinical teachers and medical education experts then developed the present 11 domains, and a further feasibility pilot study was carried out (Ker et al., 2009). A generalizability study (Ker & Till, 2014) also indicated that 11 domains would provide a reliable assessment. Assessors were trained to rate students' performance in each domain using a 5-point scale. In each of these 11 domains, the student is judged on a distinctly different performance (task) such as written documentation, communication with the patient and relatives, and so on. The assessment tool was piloted to assess the functioning of the rating instrument (McIlwaine, McAleer, & Ker, 2007). All the assessors were either trained senior clinicians with expertise as junior doctors' educational supervisors, or clinical skills educators who were familiar with the performance and assessment of senior students.

All assessors participating in the FYWSE have to undertake annual mandatory standardized training using the assessment tool (Ker et al., 2009). Each 3-hour training session covers the stages of the FYWSE and the process of carrying out the assessment. The assessors are shown a video of a student who is a clear fail and a video of one who is a clear pass. They are then asked to assess each student's performance and to summarize the behaviors observed including strengths and areas for improvement. Next, they have to assess each student on the 11 domains using the 5-point scale. The assessors' ratings are then shared and a discussion of outlier ratings facilitated. Following this part of the training, videos of one or two borderline students are shown, and the same 2-step assessment training process repeated.

To prevent students being able to prepare for one simulation exercise, the faculty developed three simulation exercises that involved different contexts and disease entities, but the elements of each exercise and the number of timed interruptions were the same across exercises.

Administration of the FYWSE. During the FYWSE simulation, each final-year student is randomly assigned to one of the three possible simulation exercises where he or she is handed over the care of six patients on the simulated ward for a 20-minute period. Students can contact a senior colleague, and they have the support of a nurse on the ward. Timed interruptions from patients and staff, as well as phone calls and pagers, are integral to the exercise, reflecting the reality of clinical practice (McIlwaine, McAleer, & Ker, 2007).

Two assessors directly observe a student's performance via a live link to a dedicated viewing room. The assessors independently rate the student, assigning separate ratings on 11 domains. Using a conjunctive approach, each assessor then makes a global pass/fail judgment, i.e., a student has to score at least a 3 on each of the 11 domains in order to pass the assessment. Stronger performances in some domains cannot compensate for a weaker performance in one of the other domains. Students who do not pass this simulation assessment are given the opportunity to take the assessment again at a later time.

After completing the assessment, each student participates in a 10-minute, face-to-face feedback session with the assessors, supported with a written summary of the student’s performance. Students’ performance on this assessment is but one source of evidence of a student’s clinical ability that is included in that student’s portfolio, alongside a variety of other reports and presentations.

Participants in the Final-Year Ward Simulation Exercise

Students. All 154 students who were in their final year of study in 2010 - 2011 participated in the FYWSE used in this analysis. Ninety-eight (64%) of the students were female and 56 (36%) male. No other demographic data such as age, or race/ethnicity of the students were available.

Assessors. The students' performance in the FYWSE was assessed by a total of 32 assessors. Each student was assessed by two assessors. Fourteen (42%) of the assessors were female and 18 (58%) male. No other demographic data such as age, or race/ethnicity of the assessors were available.

Measures and Rating Design

The FYWSE was designed to assess clinical performance on 11 core domains: (1) Task management, (2) Clinical skills, (3) Acutely ill patient, (4) Prescribing, (5) Written documentation, (6) Response to interruptions, (7) Communication with patient and relations, (8) Communication with colleagues (team), (9) Safe medical practice, (10) Health and safety (cross infection), and (11) Professionalism. Each assessor assigned one rating in each of the domains using a 5-point scale where 1 = Very Poor, 2 = Poor, 3 = Average, 4 = Good, and 5 = Outstanding. With two assessors rating them, each student received a total of 22 ratings.

Data Collection

The FYWSE took place in the dedicated simulated ward in the Clinical Skills Centre of the medical school. Three simulation exercises that involved different contexts and disease entities were created, but the elements of each exercise and the number of timed interruptions were the same across exercises. For a specific student to be assessed, the staff set up the simulated ward with a simulation exercise, and one student at a time acted as a doctor and was assessed. The students were randomly assigned to the exercises (54 students to Exercise 1, 56 to Exercise 2, and 44 to Exercise 3). Assessors were allocated to the exercises according to their availability, with one from the core clinical skills staff and one from the NHS senior clinical staff assessing each student.

Psychometric Analyses

We performed our MFRM analysis using Facets 3.65.0 (Linacre, 2010) (http://www.winsteps.com/facets.htm). We estimated variance components from a g-theory analysis on the same data set using G_String IV (Bloch & Norman, 2011). (http://fhsperd.mcmaster.ca/g_string/)/. Our purpose in running these analyses was to gather and examine evidence for a set of propositions that the designers of the FYWSE could use in building their validity argument.

Multi-facet Rasch modelling (MFRM). We conducted a MFRM analysis on the individual ratings that the assessors assigned for the 11 domains. We used a rating scale model that included four facets (i.e., students, exercises, assessors, and domains). Employing a MFRM approach to analyze the FYWSE rating data allowed us to look at individual-level effects of the various elements within each facet of our analysis (that is, how individual assessors, students, exercises, and domains performed). The unit of analysis in a MFRM analysis is the individual element within each facet.

Because the judging plan was set up so that assessors and students were nested within exercises, we needed to use a subset group anchoring procedure (Linacre, 2010) in order to create a network through which all parameters could be linked, making it possible to place all measures estimated from the ratings on one common scale (Linacre & Wright, 2002). We anchored the student subgroups (i.e., set the mean clinical ability measure of each of the three subgroups of students at zero and allowed each student's clinical ability measure to float, relative to that fixed mean). In so doing, we were making the assumption that because students were randomly assigned to the three exercises, the average clinical ability measures for those three student subgroups would be approximately equal (Linacre, 2010). Additionally, we noncentered the domains facet (rather than the students facet) so that the analysis would not be overconstrained.

The Facets computer programme estimated the clinical ability of each student and the severity of each assessor. Additionally, the programme calculated the difficulty of each exercise, the difficulty of receiving high ratings in each domain, and the difficulty of each scale category.

From our MFRM analysis we obtained measures of fit for each assessor, student, domain, and exercise. These are summary measures that describe the overall fit to the measurement model of the set of ratings associated with that particular element of the facet. For example, for an assessor, the fit measures indicate how well the ratings that the assessor assigned to all the students that he/she rated on all 11 domains fit the expectations of the measurement model. For a student, the fit measures indicate how well the ratings that the two assessors assigned to that student on the 11 domains fit the expectations of the measurement model.

Large differences between the observed and expected ratings (expressed as standardized residuals) indicate surprising or unexpected results (i.e., ratings that do not seem to “fit” with the other ratings and thus are puzzling and hard to explain). The residuals are typically summarized as mean-square (MnSq) error statistics called Outfit and Infit. The Outfit statistics are unweighted mean-squared residual statistics that are particularly sensitive to outlying unexpected ratings. By contrast, the Infit statistics are based on weighted mean-squared residual statistics and are less sensitive to outlying unexpected ratings. The MnSq value is the chi-square statistic divided by its degrees of freedom. Its expected value is 1; the range is 0 to infinity. MnSq fit values between 0.5 and 1.5 are productive for measurement, but a MnSq value less than 0.5 indicates little variation in a pattern of ratings, while a MnSq fit value greater than 1.5 indicates more than typical variation in the ratings (that is, a set of ratings with one or more unexpected or surprising ratings that don't seem to "fit" with the others).

After conducting the MFRM analyses, we identified students and assessors with Infit or Outfit MnSq statistics < 0.5 and equal to or greater than 1.5. For those students with Infit or Outfit MnSq values < 0.5, we examined the set of ratings each student received from the different assessors who judged him/her to determine whether the student's rating profile showed a lack of variability, which would indicate that the ratings were not discriminating well. For those assessors with Infit or Outfit MnSq values < 0.5, we examined the ratings they assigned to determine whether any of those assessors showed evidence of restriction of range (i.e., using only a few of the rating scale categories, and not the full scale, when evaluating students).

We also identified those students and assessors who had Infit or Outfit MnSq values equal to or greater than 1.5 and reviewed their rating profiles, identifying the particular ratings that were misfitting. It was important to determine whether any student had multiple misfitting ratings and thus exhibited an unexpected measurement pattern, or whether the misfitting ratings that assessors assigned were anomalous outliers. It was also important to determine whether there were any assessors who did not appear to be using the 5-point scale in a consistent manner when evaluating students' performances, assigning one or more surprising or unexpected ratings.

Generalizability analysis. We used the G_String IV programme to estimate variance components from a g-theory analysis on the same dataset that we analysed using Facets. As with MFRM, a g-theory analysis allows for multiple sources of error to be 'disentangled.' However, unlike MFRM, the unit of analysis is the group. A g-theory analysis begins with the G-study which estimates the variance attributed to each of the identified facets and the interactions amongst these, as well as the identification of all possible sources of error that could impact on the students' scores in that particular assessment (Brennan, 2000). The FYWSE analysis was a naturalistic design that included as facets exercise, assessor, and domain.

In the following D-study, we addressed the questions of whether it might be better to use one assessor instead of two, and whether the number of domains should be increased, or not (Streiner & Norman, 2008).

As the students completed the simulation exercises one at a time and were randomly assigned to one of three possible simulation exercises, the FYWSE was a nested design with the facet of differentiation (student) nested within the facet of stratification (exercise) (s : e). There were two facets of generalization, i.e., assessor (a), and domain (d). During the exercise, each student was assessed by two assessors. The assessors, however, also assessed other students during their assessment opportunity. The question of whether assessor was nested within student (a : s : e) or crossed with student (a x s : e) was addressed by performing two G-studies, one with student nested within assessor, and one with student crossed with assessor. This gave us the opportunity to see the effect of the design on the estimated error variances. The 11 scoring domains were common to all three simulation exercises.

In the FYWSE we want to be able to make an absolute decision about whether each student has reached the required level of clinical ability. Also, we see the three simulation exercises as a sample of a large number of such exercises that could be developed, and the assessors who judged the performance of the students as a sample of all possible assessors. In our G-studies, we thus see exercises and assessors as random rather than fixed factors, and so we report the absolute error G-coefficient (Phi coefficient (Φ)) (Streiner & Norman, 2008, pp. 215-222). Although the G-coefficient reflects the reliability of a single rating (Streiner & Norman, 2008, p. 222), we also estimated the equivalent of inter-rater reliability by setting assessor as a random effect and domain as a fixed effect, as well as the equivalent of internal consistency by setting domain as a random effect and assessor as a fixed effect.

We conducted a D-study for the assessor-nested-in-student design, investigating whether a decrease in the number of assessors (i.e., only one assessor judging each student), or an increase or decrease in the number of domains (i.e., more or less than 11) on which the students were rated, would have the effect of increasing the overall reliability.

Ethical Approval

The manuscript was submitted to the University of Dundee Research Ethics Committee (UREC). It met the ethical standards, but did not require ethical approval.

Results

Multi-facet Rasch Modelling

We ran a MFRM analysis on the 3,327 ratings from the assessment. There were 61 (1.8%) missing ratings (missing at random mostly due to an assessor returning a “n/a” comment instead of a rating), and they were not imputed. Facets reported that there was sufficient linkage between the facets under investigation so that they could thus be calibrated onto a single linear logit scale (Figure 1).

Global fit statistics. The Rasch measures accounted for 58.06 % of the total variance in the ratings. The variance explained by the students was 48.6%; by the assessors, 4.3%; by the exercises, 0.4%; and by the domains, 4.8%.

Student performance. The first column in the variable map shown in Figure 1 displays the equal-interval logit (log-odds) scale, ranging from -5 to +6 logits, upon which all facets analyzed were positioned. Column 2 displays the student clinical ability measures ordered from higher performing at the top (highest performance measure = 5.03 logits) to lower performing at the bottom (lowest performance measure = -4.35 logits) with a range of 9.38 logits. An asterisk denotes two students and a dot one student. The FYWSE succeeded in reliably (.96) separating the students into about seven statistically distinct levels of clinical ability.

When we inspected the fit statistics for individual students (Table 2), the majority of students (114 or 74.0%) had Infit and Outfit MnSq values between 0.5 and 1.5, which indicated that the ratings they received were productive for measurement. Sixteen (10.4%) of the students had Infit or Outfit MnSq statistics < 0.5. An inspection of the ratings given to each of these students showed that they all exhibited some variability in their rating profiles; thus, none of the 16 student rating profiles suggested a lack of precision in the measurement of the 11 domains that were assessed. However, inspection of the rating profiles of these 16 students did show a high amount of agreement (i.e., 84.5%) in the ratings that the two assessors who judged each student assigned for each domain.

Table 2 also shows that 15 (9.7%) of the students had Infit MnSq and 14 (9.1%) had Outfit MnSq values equal to or greater than 1.5 (but less than 2.0), and 9 (5.8%) had Infit and Outfit MnSq values greater than 2.0. These results suggest that one or more ratings that assessors assigned to each of these students were somewhat surprising or unexpected, given the other ratings that each student received (Wright & Linacre, 1994). This indicates that the rating profiles for these students showed an unacceptable amount of variability.

Assessor performance. Column 3 of Figure 1 displays the assessor severity measures ordered from the most stringent assessor at the top (1.02 logits) to the least stringent assessor at the bottom (-2.07 logits), a range of 3.09 logits. Each asterisk denotes one assessor. The results presented in Figure 1 indicate that there were differences between assessors in their levels of severity/leniency. The assessor separation index, which is a measure of the number of statistically distinct levels (or strata) of assessor severity in a sample of assessors (Wright & Masters, 1982), was 3.12 with a separation reliability of .81. These results suggest that there were about three strata of severity levels, showing that the assessors did not all rate in an interchangeable manner.

When we inspected the fit statistics for individual assessors (Table 2), a total of 3 (9.4%) of the 32 assessors had Infit and Outfit MnSq values equal to or greater than 1.5, signifying that these three assessors assigned one or more ratings to students that were surprising or unexpected, given how each assessor used the rating scales to rate other students. Sixteen (50.0%) of the assessors had Infit MnSq values less than 1.0 and 18 (56.2%) had Outfit MnSq values less than 1.0 signifying a high percentage of agreement between ratings that these assessors assigned. This result is supported by the fact that the assessors' ratings were in exact agreement 64% of the time.

Exercise performance. Students were allocated randomly to one of the three exercises with 54 students taking Exercise 1, 56 taking Exercise 2, and 44 taking Exercise 3. Column 4 of Figure 1 shows the difficulty levels for the three exercises, ordered from most to least difficult to obtain high ratings on. Exercise 1 was the most difficult exercise (0.21 logits, SE 0.05) while Exercise 3 was the easiest (-0.39 logits, SE 0.05), a range of 0.60 logits. Average scores were 72.19%, 73.34%, and 75.93% for Exercises 1, 2 and 3 respectively. We ran a one-way ANOVA on the three exercise difficulty measures and then performed post-hoc comparison tests. The results indicated that the measures for Exercises 1 and 3 were statistically significantly different from one another, as were the measures for Exercises 2 and 3, but the measures for Exercises 1 and 2 were not significantly different from one another ($\alpha = 0.05$). (The post-hoc comparison tests we performed made use of the Bonferroni correction to adjust for multiple comparisons.)

All three of the exercises had Infit and Outfit MnSq values between 0.96 and 1.03, which suggests that the ratings that the assessors assigned to the exercises were productive for measurement.

Domain performance. Column 5 of Figure 1 shows the difficulty levels for the 11 domains upon which all students were assessed. The domains had differing difficulty levels (i.e., some were more difficult to obtain high ratings on than others), ordered from the most difficult at the top (Domain 10, Health and Safety – Cross Infection) at 0.07 logits (SE 0.09) to the least difficult at the bottom (Domain 11, Professionalism) at -1.70 logits (SE 0.10), a range of 1.77 logits. The domain separation index, which is a measure of the number of statistically distinct levels (or strata) of difficulty in a sample of domains (Wright & Masters, 1982), was 8.26 with a reliability of .97.

All but one of the domains had Infit and Outfit MnSq values between 0.5 and 1.5, which indicates that the ratings the assessors assigned to the various domains were productive for measurement. The one exception was Domain 10, Health and Safety - Cross Infection, which had Infit and Outfit MnSq values of 1.93, indicating that some of the ratings that the assessors assigned for that domain were quite surprising and unexpected, given the other ratings that students received. These results suggest that perhaps this domain is measuring an aspect of clinical ability that may be qualitatively different from what the other domains are measuring. The assessors seem to be using the 5-point scale in a somewhat different manner when rating students in this particular domain, implying that perhaps the FYWSE is functioning as a multidimensional instrument.

Rating scale performance. Column 6 of Figure 1 shows the 5-point rating scale that the assessors used when rating students. The horizontal lines show the scale category thresholds. (A category threshold denotes the point at which the likelihood of the student receiving the next higher rating is equal to the likelihood of that student receiving the next lower rating.) For example, students with clinical ability measures between 0.93 and 3.97 logits were more likely to receive a rating of 4 (Good) than any other rating, while students with clinical ability measures above 3.97 logits were more likely to receive a rating of 5 (Outstanding) than any other rating. The student with the lowest clinical ability measure of -3.57 logits was most likely to receive a rating of 2 (Poor).

Table 3 presents information regarding the functioning of the 5-point rating scale indicating that the assessors used the middle scale categories – i.e., 3 (Average) and 4 (Good) most frequently at 34% and 38% of the time, respectively. The third and fourth columns in Table 3 report the average measures and the Rasch-Andrich thresholds for each rating category. If the rating scale were working as intended, we would expect that the average measures and the Rasch-Andrich thresholds for the categories should increase as the rating scale categories increase, i.e., higher clinical ability should correspond to higher average measures and higher Rasch-Andrich thresholds. As shown in Table 3, the average measures and the Rasch-Andrich thresholds all increased in value. Additionally, apart from the difference between categories 1 and 2 where the difference is 0.93, the average measures advanced by more than 1.4 logits (Linacre, 2002). Taken together, these findings suggest that the 5-point rating scale functioned as those who designed it intended, and that students with higher ratings exhibited more of the construct being measured (i.e., clinical ability) than did students with lower ratings. The last column in Table 3 shows the Outfit Mean-Square (MnSq) statistic for each rating category. The values shown are all close to the expected value of 1.0. This finding suggests that for each rating category, the average student clinical ability measure is close to the measure the Facets model would predict for that rating category if the data were to fit the model perfectly, which provides additional evidence to support the claim that the 5-point rating scale functioned as intended (Linacre, 1999).

Diagnosing misfit. In this FYWSE, 148 (4.4%) of the 3,327 ratings that assessors assigned had standardized residuals equal to (or greater than) + or - 2, suggesting that those particular ratings were unexpected and did not fit the measurement model. Even though overall the data showed good fit to the model, with the total number of unexpected ratings well within the accepted norms, we investigated the unexpected ratings by student, assessor, exercise, and domain.

Eighty (51.95%) of the 154 students were assigned 1 to 6 unexpected ratings each. The majority of these students had between 1 and 4 unexpected ratings, with two students having 5 and only one student having 6 unexpected ratings.

Assessors assigned unexpected ratings when evaluating students’ performances in all three exercises. They assigned 46 (31.1%) unexpected ratings when assessing students in Exercise 1, 55 (37.2%) in Exercise 2, and 47 (31.8%) in Exercise 3.

Assessors also assigned unexpected ratings in all 11 of the Domains. They assigned 3 (2.0%) of those 148 unexpected ratings in Domain 8, Communication with Team, and between 7 (4.7%) and 17 (11.5%) unexpected ratings in most of the other domains. However, the assessors assigned the most unexpected ratings (i.e., 48 or 32.4%) in Domain 10, Health and Safety - Cross Infection. There were 19 different assessors who were responsible for assigning those 48 unexpected ratings. They involved 39 different students who participated in all three exercises.

The vast majority of the assessors (29 or 90.6%) assigned at least some unexpected ratings, i.e., each of these assessors assigned between 1 and 14 unexpected ratings. Only four of the assessors assigned no unexpected ratings. Inspection of the unexpected ratings of those assessors who had assigned the majority of unexpected ratings (7 or more), indicated that the 104 unexpected ratings they assigned were also spread over all three exercises with 29 (27.9%) occurring in Exercise 1, 36 (34.6%) in Exercise 2, and 39 (37.5%) in Exercise 3.**Impact of differences in assessor severity and exercise difficulty on student clinical ability measures**. The Facets computer programme adjusted the student clinical ability measures for differences in the levels of severity/leniency that individual assessors exercised and for differences in the difficulty of obtaining high ratings in the different exercises. Since different sets of two assessors rated each student, some students may have been unfairly advantaged if they happened to be rated by more lenient assessors, while other students may have been unfairly disadvantaged if they happened to be rated by more stringent assessors. Additionally, since each student participated in one of the three exercises, some students may have been unfairly advantaged if they happened to be assigned to an easier exercise, while other students may have been unfairly disadvantaged if they happened to be assigned to a more difficult exercise. For each student, Facets calculated a "fair average" measure, which indicated the student's level of clinical performance had assessors of average severity/leniency rated that student and had the exercises been of equal difficulty (Linacre, 1997).

Generalizability Theory

In the G-studies we simultaneously estimated the variance components associated with the logistical stratification facet (exercise (e)), as well as the other two major sources of measurement error, the facets of generalization (i.e., assessor (a), and domain (d)) impacting on the universe scores of the facet of discrimination – the 154 students (s) who completed the FYWSE. There were 61 (1.8%) missing ratings (missing at random mostly due to an assessor returning a "n/a" comment instead of a rating), which we replaced with the average rating of 3. The replacement of the missing ratings with the average was performed due to the inability of g-theory, in contrast with MFRM, to deal with missing data.

As only one student at a time could perform the simulation exercise, each student was rated by only two assessors; thus, assessor was nested within student (a : s : e), which suggests a g-theory analysis would be appropriate. The estimated variance components for each source of variation in this design are shown in Table 4. As we wished to place an absolute interpretation on the outcome of the FYWSE, variance due to strata (exercise) is a source of error that is included in the Absolute Error calculation. The G-coefficient (Φ) for this analysis was 0.79, the equivalent of an inter-rater reliability of .89, and the equivalent of an internal consistency index of .88. Each assessor rated more than one student, which suggests that a crossed g-theory analysis (a x s : e) might also be appropriate. Table 5 shows the estimated variance components for each source of variation in this design. It must be remembered that the estimated variance components shown in Tables 4 and 5 are for single observations only (Streiner & Norman, 2008, p.222). From these single-observation variance components, we identified and addressed sources of variation that contribute the most error.

The goal of a psychometric analysis is to produce measures that differentiate between the subjects of observation (students) (Streiner & Norman, 2008, p.225), and thus one would prefer that most of the estimated variance be attributable to students. Tables 4 and 5 indicate that, as was the case with the results from the MFRM analysis, the major portion of the variance component (38%) was attributed to students (s : e), suggesting that the students differed considerably in their demonstrated clinical ability. The other variance components listed all represent sources of measurement error. The variance component attributed to the facet of stratification, exercise (e), on its own accounted for only 1.7% of the total score variance, showing that there were not large differences in the levels of difficulty between simulation exercises. This is a promising finding. If the exercises differed in their levels of difficulty, that would indicate bias in the assessment. Assessors on their own (Table 5) accounted for only 0.59% of the total score variance, while assessors nested in students (a : s : e) accounted for only 3.3% of the total score variance. This indicates that the assessors showed considerable agreement in the ratings that they assigned, suggesting that the assessors rank ordered the students similarly. This result was supported by the previously reported inter-rater reliability of .89.

The next largest variance components include the variances attributed to domain, which collectively accounted for 57% of the total score variance (inconsistencies) in the ratings. This could be an indication that it was more difficult to be rated highly on some of the domains, and that the rank order of the students might not have been the same on all of the domains.

In the D-study we investigated the projected reliability when decreasing the number of assessors (i.e., having only one assessor rate each student), or increasing or decreasing the number of domains on which the students were rated. Table6 shows that the reliability of the assessment results increases slightly with an increased number of assessors and rating opportunities (domains).

Discussion

The purpose of the FYWSE as used at Dundee Medical School is to give senior students the opportunity to care for patients in an authentic simulated clinical setting while also providing the institution the opportunity to assess whether the students have the necessary level of clinical ability to provide high quality care for their future patients. Faculty who designed the assessment intend for the final judgment to represent an individual senior medical student's level of clinical performance at that point in time.

We now examine the evidence for each proposition that is part of the validity argument for the FYWSE to (1) evaluate the soundness of the proposition, and (2) determine to what extent the evidence we reported appears to support the proposed score interpretation (American Educational Research Association et al., 2014). We conclude by considering the implications of our research for the ongoing FYWSE validation process.

Proposition 1. The FYWSE measures critical performance behaviours that are required to provide high quality patient care in the context of the NHS.

The faculty who designed the FYWSE ensured that the simulation exercises used to assess the clinical performance of the senior students were authentic and reflected the reality of patient care by an iterative process, first getting experts in the field to identify the actual clinical performance behaviours required in the workplace in which the students will have to function as junior doctors, and then by mapping these performance behaviours against the Tomorrow's Doctors (General Medical Council, 2009) and the Foundation Programme curriculum outcomes (Foundation Programme, 2012). The process used to develop the three main elements (a new admission, communication issues, and the management of the acutely unwell patient) that make up each simulation exercise also included focus groups of health care practitioners who work with junior doctors, reviewing admissions to an acute medical unit, reviewing commonly reported errors, and observational shadowing of junior doctors to identify common interruptions and task management issues. A literature review on both the workload of junior doctors and common admissions was also undertaken.

In the simulated exercises, the authentic context that reflects the complex real-world pressures that junior doctors face in the workplace is created by using trained simulated patients who intentionally interrupt the doctor. Health care practitioners are also scripted to interrupt through the pager system, by telephone, or personally.

Judgment: The proposition seems sound, and the evidence gathered appears to support the proposed score interpretation. This assessment consists of the systematic observation of a student's behaviour in an authentic simulated situation where the content of the assessment has been carefully aligned with prerequisite performance behaviours identified through various means. Construct underrepresentation has been avoided through the inclusive process used to develop the three main elements of the exercises, and the dedicated simulated ward has minimized the effect that construct irrelevance could have on the interpretation of scores. Although expensive and time-consuming, the faculty should consider increasing the number of exercises used for this purpose in order to better represent the content domain, especially if they intend to use this assessment with other groups or for high-stakes assessments that could have career-defining consequences.

Proposition 2. Students who perform better on the FYWSE will be more able to provide high quality patient care than those who do not perform well.

The assessment was built around the identified elements related to patient care, i.e., a new admission, communication issues, and the management of the acutely unwell patient. The simulation exercises created provided an authentic context for the assessment. The assessment designers ensured that each of the three simulation exercises contained the three main elements.

The data we used in our analyses demonstrated acceptable fit to the MFRM model, and our results showed that the FYWSE could reliably (.96) separate the students into distinct levels of clinical ability, giving us reassurance that the FYWSE was able to identify those students who are competent in the clinical setting, as well as those who still needed some practice.

Although their methods for estimating the variance components are different, the results from the g-theory and MFRM analyses both indicated that the largest proportion of the variability in the assessment was accounted for by differences in students' clinical performance. A large amount of the variance in ratings thus appears to be attributable to systematic differences in the clinical performance of the students.

Judgment: The proposition seems sound, and the evidence gathered appears to support the proposed score interpretation. However, the medical school faculty might consider obtaining additional concurrent and/or predictive validity evidence in order to determine whether those students who performed well on this assessment do, indeed, provide high quality patient care in other settings.

Proposition 3: Students responded to the exercises in the manner that the assessment designers intended.

The results from the MFRM analysis indicated that the vast majority of the students (i.e., 130 or 85%) appeared to have responded to the exercises in the manner that the assessment designers intended, showing little evidence of inconsistencies in their rating patterns. However, for the remaining 24 students (i.e., 15%), the assessors assigned one (or more) ratings that were somewhat surprising or unexpected, given the other ratings that each student received. That is, those particular ratings did not seem to "fit" with the other ratings and thus were puzzling and hard to explain.

Judgment: The proposition seems sound; and, for the vast majority of the students, the evidence gathered appears to support the proposed score interpretation. Currently, students' scores on this assessment are used as one piece of evidence of their clinical ability, alongside a variety of other timed reports and presentations in their portfolios. However, in the future, if the medical school faculty were to decide to use students' scores on the FYWSE for higher stakes summative decision making, then they might want to consider instituting a quality control monitoring procedure to review the rating profiles of those students who are rated inconsistently before releasing score reports to help ensure greater fairness in the application of their assessment process.

Proposition 4: Assessors who judged the students were able to evaluate the students' clinical performances appropriately and fairly.

All the assessors were either trained senior clinicians with expertise as junior doctors' educational supervisors, or clinical skills educators who were familiar with the performance and assessment of senior students. The assessors undergo annual standardized training.

The assessors' ratings were in exact agreement 64% of the time, and their interrater reliability was .89. According to the results from our g-theory analyses, the amount of total score variance attributed to assessors was small (i.e., less than 4%), suggesting that the assessors rank ordered the students similarly. Our results are encouraging, as they suggest that for the FYWSE, differences in assessor severity are not as pronounced as they are in other types of performance assessments such as MMIs (Roberts, Rothnie, Zoanetti, & Crossley, 2010; Till, Myford, & Dowell, 2013) and clinical assessments (McManus, Thompson, & Mollon, 2006).

However, while there was a considerable amount of agreement among the assessors in the ratings they assigned, the results from our MFRM analysis indicated that the assessors were not interchangeable. Among the 32 assessors who participated in this assessment, there were three statistically distinct strata of severity levels represented, which may have led to some unfairness in the final outcomes for some students. When assessors are not interchangeable, then a student's score will depend not only on the quality of his/her clinical performance, but also in part on which particular assessors happened to assign the ratings.

The results from the MFRM analysis indicated that the vast majority of the assessors (i.e., 29 or 90.6%) assigned ratings in a consistent manner, using the rating scale as the assessment designers intended. However, the remaining 3 assessors (i.e., 9.4%) assigned some ratings that were somewhat surprising or unexpected.

Judgment: The proposition seems sound for the majority of the assessors; and, by and large, the evidence gathered appears to support the proposed score interpretation. However, if the medical school faculty were to decide to use the students' scores for higher stakes summative decision making, then before issuing final score reports for students, it would be prudent for faculty to review unexpected ratings to determine whether they seem warranted, or not. Faculty might also want to consider adjusting students' scores for the impact of differences in assessor severity in order to improve the fairness of the assessment process.

Proposition 5: The instrument used to assess the students' clinical ability during the FYWSE functioned as intended.

This proposition requires discussion of evidence presented for exercises, domains, and rating scales:

Exercises. Students were randomly assigned to complete one of three possible simulation exercises, but unfortunately, the use of random assignment did not succeed in creating a level playing field for all students. The results from our g-theory analyses revealed that the variance component attributed to the exercises accounted for only 1.7% of the total scores variance, showing that there were not large differences in the levels of difficulty between simulation exercises. However, while those differences were not large, the results from the MFRM analysis indicated that Exercises 1 and 2 were significantly harder to get high ratings on than Exercise 3. Average scores were 72.19%, 73.34%, and 75.93% for Exercises 1, 2 and 3 respectively. Differences in difficulty could impact students' scores. Students who were assigned to Exercises 1 or 2 were at more of a disadvantage than students who were assigned to Exercise 3. That is, it was harder to do well on Exercises 1 and 2 than on Exercise 3.

There are several possible explanations for these findings. There may be aspects of Exercise 3 that make it easier for students to obtain high ratings. Another possible explanation might be that the assessors who were assigned to evaluate students' performances for Exercise 3 were more lenient than the assessors who evaluated students' performances for Exercises 1 and 2, since unlike the students, the assessors were not randomly assigned to the exercises. From our findings, it is difficult to determine whether Exercise 3 was less demanding, or whether the assessors who were assigned to that exercise tended to assign higher ratings on average than did the assessors who were assigned to the other two exercises, which would have made Exercise 3 appear easier. The nesting of assessors and students within exercises in this judging plan makes it impossible for us to disentangle the confounded effects of assessor severity and exercise difficulty. A third possible explanation for these findings might be that the students who were assigned to Exercise 3 tended to be the higher ability students. However, since students were randomly assigned to the exercises, it doesn't seem as though that should be the case. Random assignment should have resulted in a fairly equal distribution of student abilities across the three exercises.

Domains. Among the 11 domains, Domain 11, Professionalism, was the easiest to score high ratings on, and Domain 10, Health and Safety – Cross infection, was the most difficult. All but one domain (Domain 10, Health and Safety – Cross Infection) had acceptable fit statistics. The results from the MFRM analysis indicate that students found this the most difficult domain to score high ratings on, and the ratings that students received for this domain showed low correlations with their total scores. This domain had an Infit MnSq value of 1.93, and assessors assigned most (48%) of the unexpected ratings when evaluating students in this domain. As these unexpected ratings were assigned by 19 (58%) of the assessors to 39 different students in all three of the exercises, it seems quite evident that the assessors are using the 5-point rating scale in a different manner when they were rating student performance in this particular domain. Furthermore, the finding that scores for this domain show a low correlation with students' total scores suggests that the ratings that the assessors assigned for this domain seemed to be measuring aspects of students' clinical performance that are qualitatively different from, and quite unrelated to, those aspects of students' clinical performance that the other domains are measuring, perhaps suggesting that the FYWSE may be functioning as a multidimensional instrument.

The large variance components attributed to domain in the generalizability analyses which collectively account for 57% of the total variance (inconsistencies) in the ratings could be an indication that it was more difficult to be rated highly on some of the domains, and that the rank order of the students might not have been the same on all of the domains. These findings support the MFRM results.

Rating scale. The results from our MFRM analysis indicated that the 5-point scale used to rate the students functioned as those who designed it intended, and that students with higher ratings exhibited more of the construct being measured (i.e., clinical ability) than did students with lower ratings (Linacre, 1999).

Judgment: The proposition seems sound; and, for the most part, the evidence gathered appears to support the proposed score interpretation. However, if the medical school faculty should decide to use the students' scores for higher stakes decision making, they might consider adjusting student scores for the impact of differences in exercise difficulty to improve the fairness of the assessment process. Additionally, the assessment designers may want to consider whether it is advisable to continue to include Domain 10, Health and Safety – Cross Infection, in the assessment. If they decide that this domain is critical to the measurement of students' clinical ability, then there may be a need for some revisions of the domain description, the rating scale, and/or the training of the assessors.

Proposition 6: Eleven domains and two assessors are sufficient to produce dependable, reliable/precise scores across replications of this assessment procedure.

The results from both the MFRM and g-theory analyses indicated that the scores obtained from the FYWSE across the different replications of the simulation exercises were reliable. The MFRM results showed that the FYWSE succeeded in reliably (.96) separating the students according to levels of clinical ability, and the results from the G-studies showed a G-coefficient of 0.79 and an equivalent of an internal consistency index of .88.

The results from subsequent D-studies investigating the projected reliability when increasing or decreasing the number of assessors and/or domains indicated that the generalizability coefficient with 11 domains would be .80 for one assessor and .84 for two assessors.

Judgment: The proposition seems sound, and the evidence gathered appears to support the proposed score interpretation. The results indicate that the assessment succeeded in reliably differentiating between the students on the basis of their demonstrated clinical ability, and that the current use of two assessors and 11 domains produces reliable scores.

Conclusions

Our purpose in conducting this study was to illustrate how assessment designers who are creating assessments to evaluate medical students' clinical performance might approach the tasks of developing propositions and collecting and examining various sources of evidence to build a validity argument. We used as our example the final-year Ward Simulation Exercise (FYWSE), created by faculty and the University of Dundee Medical School to assist them in determining whether their senior medical students have acquired the level of clinical ability needed to provide high quality patient care.

Using multi-facet Rasch measurement and generalizability theory approaches, we examined various sources of validity evidence that the medical school faculty have gathered for a set of six propositions needed to support their use of scores in this manner. Based on our analysis of the evidence, we would conclude that the propositions appear to be sound, and that the evidence they have gathered seems to support their proposed score interpretation. Given the body of validity evidence they have collected thus far, their intended interpretation seems defensible.

We would encourage the medical school faculty to continue to build their validity argument, identifying other sources of evidence that they can present to further support these propositions. Additionally, the faculty might consider other propositions that are needed to bolster their validity argument, and then identify sources of validity evidence for those propositions. To assist them in that task, they might consider rival hypotheses that could serve to challenge their proposed score interpretation. They may also find it useful to contemplate potential unintended consequences of FYWSE use, as well as additional possible sources of construct-irrelevant variance that might affect students’ scores. Finally, as they are reviewing the 11 domains that are currently included in this assessment, faculty might also consider whether there are any important aspects of the clinical ability construct that the FYWSE fails to capture that might distort the meaning of the scores (i.e., construct underrepresentation) (American Educational Research Association et al., 2014, pp. 12-13).

The assessment designers carefully aligned the content of the simulation exercises used in the FYWSE with prerequisite performance behaviours for junior doctors, and they used the General Medical Council's good medical practice to construct the domains. A logical extension of this simulation assessment could thus be to use it to diagnose areas of concern for doctors after they have started their posts in the clinical workplace. If the assessment designers were to decide to use the FYWSE for this new purpose (or for any other purposes that carry higher stakes than the current purpose), then the character of the evidence needed would change (American Educational Research Association et al., 2014, p. 22). Building validity arguments to support those new uses would require the gathering of additional validity evidence, with higher standards attached to that evidence.

Acknowledgements, Funding, Disclaimers, Ethical Approval

1. *Acknowledgements:* None
2. *Funding/ Support*: None
3. *Other disclosures*: None
4. *Ethical Approval*: The manuscript was submitted to the University of Dundee Research Ethics Committee (UREC). It met the ethical standards, but did not require ethical approval.
5. *Disclaime*r: None
6. *Previous presentations*: None

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *The standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Barsuk, J., Cohen, E., McGaghie, W., & Wayne, D. (2010). Long-term retention of central venous catheter insertion skills after simulation-based mastery learning. *Academic Medicine*, *85* (10 Suppl), S9-12.

Bindal, T., Wall, D., & Goodyear, H. M. (2011). Trainee doctors' views on workplace-based assessments: Are they just a tick box exercise? *Medical Teacher*, *33*(11), 919-927. doi: 10.3109/0142159X.2011.558140

Bloch, R. & Norman, G. (2011). *G String IV User Manual*. Retrieved from http://fhsperd.mcmaster.ca/g_string/download/g_string_4_manual_611.pdf

Brennan, R. L. (2000). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement*, *24*(4), 339-353.

Cook, D. A., Brydges, R., Zendejas, B., Hamstra, S. J., & Hatala, R. (2013). Technology enhanced simulation to assess health professionals: A systematic review of validity evidence, research methods and reporting quality. *Academic Medicine*, *88*(6), 873-83.

Crofts, J. F., Ellis, D., Draycott, T. J., Winter, C., Hunt, L. P., & Akande, V. A. (2007). Change in knowledge of midwives and obstetricians following obstetric emergency training: A randomized controlled trial of local hospital, simulation centre and teamwork training. *British Journal of Obstetrics and Gynaecology, 114*(12), 1534-1541.

Eckes, T. (2011). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*. Frankfurt am Main, Germany: Peter Lang.

Epstein, R. (2007). Assessment in medical education. *The New England Journal of Medicine*, *356*, 387-96.

Foundation Programme. (2012, July). *The UK Foundation Programme Curriculum Updated for August 2014*. Retrieved from http://www.foundationprogramme.nhs.uk/pages/home/curriculum-and-assessment/curriculum2012

General Medical Council. (2009). *Tomorrows doctors: Outcomes and standards for undergraduate medical education*. London, UK: Author. Retrieved from http://www.gmc-uk.org/static/documents/content/Tomorrows_Doctors_0414.pdf

General Medical Council. (2010). *Standards for curricula and assessment systems.* London, UK: Author. Retrieved from http://www.gmc-uk.org/Standards_for_curricula_and_assessment_systems_0414.pdf_48904896.pdf

General Medical Council. (2011). *Assessment in undergraduate medical education. Advice supplementary to Tomorrow's Doctors* (2009). London, UK: Author. Retrieved from http://www.gmc-uk.org/static/documents/content/Assessment_in_undergraduate_medical_education_0211.pdf

Grantcharov, T. P., Kristiansen, V. B., Bendix, J., Bardram, L., Rosenberg, J., & Funch-Jensen, P. (2004). Randomized clinical trial of virtual reality simulation for laparoscopic skills training. *British Journal of Surgery*, *91*(2), 146-50.

Hayes, K. (2011). Work-place based assessment. *Obstetrics, Gynaecology, & Reproductive Medicine, 21*(2), 52-54.

Iramaneerat, C., Yudkowsky, R., Myford, C. M., & Downing, S. M. (2008). Quality control of an OSCE using generalizability theory and many-faceted Rasch measurement. *Advances in Health Sciences Education: Theory and Practice, 13*(4), 479-493. doi:10.1007/s10459-007-9060-8

Issenberg, S. B., & Scalese, R. J. (2007). Best evidence on high fidelity simulation: What clinical teachers need to know. *The Clinical Teacher*, *4*(2), 73-77. doi: 10.1111/j.1743-498X.2007.00161.x

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). Westport, CT: Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1-73. doi: 10.1111/jedm.12000

Ker, J., & Bradley, P. (2014) Simulation in medical education. In T. Swanwick (Ed.), *Understanding medical education: Evidence, theory and practice* (2nd ed., pp. 175-192). Hoboken, NJ: Wiley-Blackwell.

Ker, J., Hesketh, A., Anderson, F., & Johnston, D. (2005). PRHO views of the usefulness of a pilot ward simulation exercise. *Hospital Medicine*, *66*(3), 168-70.

Ker, J. S, Hesketh, E. A., Anderson, F., & Johnston, D. A. (2006). Can a ward simulation exercise achieve the realism that reflects the complexity of everyday practice junior doctors encounter? *Medical Teacher*, *28*(4), 330-334. doi: 10.1080/01421590600627623.

Ker, J., Murphy, D., Anderson, F., Hogg, G., Hesketh, A., Hanslip, J., Kellett, C., & Stirling, K. (2009, July). *Reliability of a diagnostic tool to assess performance of foundation doctors in a ward simulation exercise.* Poster presented at the annual scientific meeting of the Association for the Study of Medical Education (ASME). Edinburgh, UK.

Ker, J., & Till, H. (2014). Psychometrics. In P. Dasgupta, K. Ahmed, P. Jaye, & M. Khan (Eds.), *Surgical Simulation* (pp. 95-109). Anthem Press.

Linacre, J. M. (1996). Generalizability theory and many-facet Rasch measurement. In G. Engelhard & M. Wilson (Eds.), *Objective measurement: Theory into practice* (Vol. 3, pp. 85-98). Norwood, NJ: Ablex.

Linacre, J. M. (1997). Communicating examinee measures as expected ratings. *Rasch Measurement Transactions, 111*(1), 550-551. Retrieved from http://www.rasch.org/rmt/rmt111m.htm

Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement*, *3*, 103-122.

Linacre, J. M. (2001). Generalizability theory and Rasch measurement. *Rasch Measurement Transactions*, *15*, 806-807.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*(1), 85-106.

Linacre, J. M. (2010). FACETS (Version 3.67.1)[Computer software]. Minnetonka, MN: SWReg Digital River, Inc.

Linacre, J. M., & Wright, B. D. (2002). Construction of measures from many-facet data. *Journal of Applied Measurement*, *3*(4), 486-512.

Linstone, H., & Turoff, M. (1975). (Eds). *The Delphi method: Techniques and applications*. Reading, MA: Addison-Wesley.

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing*, *15*, 158-180.

Maran, N. J., & Glavin, R. J. (2007). Low-to-high-fidelity simulation: What clinical teachers need to know. *The Clinical Teacher*, *4*, 73-77.

McIlwaine, L. M., McAleer, J. P. G., & Ker, J. S. (2007). Assessment of final year medical students in a simulated ward: Developing content validity for an assessment instrument. *International Journal of Clinical Skills*, *1*(1), 33-35.

McKinley, R. K., Strand, J. Ward, L., Gray, T., Alun-Jones, T., & Miller, H. (2008) Checklists for assessment and certification of clinical procedural skills omit essential competencies: A systematic review. *Medical Education, 42*(4), 338-349.

McLeod, R., Mires, G. J., & Ker, J. (2012). Direct observed procedural skills assessment in the undergraduate setting. *The Clinical Teacher, 9*(4), 228-32. doi: 10.1111/j.1743-498X.2012.00582.x

McManus, I. C., Thompson, M., & Mollon, J. (2006). Assessment of examiner leniency and stringency ('hawk-dove effect') in the MRCP(UK) clinical examination (PACES) using multi-facet Rasch modelling. *BMC Medical Education, 6*, 42. doi:10.1186/1472-6920-6-42. Retrieved from http://www.biomedcentral.com/1472-6920/6/42

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York, NY: Macmillan

Messick, S. (1996). Validity of performance assessment. In G. W. Phillips (Ed.), *Technical issues in large-scale performance assessment* (pp. 1-18). Washington, DC: U. S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics.

Miller, A., & Archer, J. (2010). Impact of workplace based assessment on doctors' education and performance: A systematic review. *British Medical Journal, 341*, c5064. doi: 10.1136/bmj.c5064

Miller, D. M., & Linn, R. L. (2000). Validation of performance-based assessments. *Applied Psychological Measurement*, *24*(4), 367-378. doi: 10.1177/01466210022031813

Morris, M. C., Gallagher, T. K., & Ridgway, P. F. (2012). Tools used to assess medical students competence in procedural skills at the end of a primary medical degree: A systematic review. *Medical Education Online, 17*. doi: 10.3402/meo.v17i0.18398. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3427596/

Murphy, D., Bruce, D., Mercer, S., & Eva, K. W. (2009). The reliability of workplace-based assessment in postgraduate medical education and training: A national evaluation in general practice in the United Kingdom. *Advances in Health Sciences Education: Theory and practice, 14*(2), 219-232. doi: 10.1007/s10459-008-9104-8

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, *4,* 386-422.

Nestel, D., Groom, J., Eikeland-Husebo, S., & O'Donnell, S. M. (2011). Simulation for learning and teaching procedural skills. *Simulation in Health Care, 6*, S10-13.

Norcini, J. J. (2003). ABC of learning and teaching in medicine: Work based assessment. *British Medical Journal*, *326*(7392), 753-755. doi: 10.1136/bmj.326.7392.753

Norcini J. J. (2005). Current perspectives in assessment: The assessment of performance at work. *Medical Education*, *39*(9), 880-889. doi: 10.1111/j.1365-2929.2005.02182.x

Norcini J. J., & McKinley, D. W. (2007). Assessment methods in medical education. *Teaching and Teacher Education*, *23*, 239-250.

Postgraduate Medical Education and Training Board and Academy of Medical Royal Colleges. (2009). *Workplace based assessment: A guide for implementation*. London, UK: Author. Retrieved from http://www.gmc-uk.org/Workplace_Based_Assessment___A_guide_for_implementation_0410.pdf_48905168.pdf

Roberts, C., Rothnie, I., Zoanetti, N., & Crossley, J. (2010). Should candidate scores be adjusted for interviewer stringency or leniency in the multiple mini-interview? *Medical Education*, *44*(7), 690-698. doi: 10.1111/j.1365-2923.2010.03689.x

Schuwirth, L. W. T., & van der Vleuten, C. P. M. (2003). The use of clinical simulations in assessment. *Medical Education*, *37*(Suppl 1), 65-71. doi: 10.1046/j.1365-2923.37.s1.8.x

Siassakos, D., Bristowe, K., Draycott, T. J., Angouri, J., Hambly, H., Winter, C., Crofts, J. F., Hunt, L. P., & Fox, R. (2011). Clinical efficiency in a simulated emergency and relationship to team behaviours: A multisite cross-sectional study. *BJOG: An International Journal of Obstetrics and Gynaecology*, *118*(5), 596-607. doi: 10.1111/j.1471-0528.2010.02843.x

Smith, E. V., & Kulikowich, J. M. (2004). An application of generalizability theory and many-facet Rasch measurement using a complex problem-solving assessment. *Educational and Psychological Measurement, 64*, 617-639. doi: 10.1177/0013164404263876

Stiggins R. (2005). From formative assessment to assessment FOR learning: A path to success in standards-based schools. *Phi Delta Kappan*, *87*(4), 324-328.

Stiggins, R., & Chappuis, J. (2006). What a difference a word makes. *Journal of Staff Development*, *27*(1), 10-14.

Streiner, D. L., & Norman, G. R. (2008). *Health measurement scales: A practical guide to their development and use* (4th ed.). New York, NY: Oxford University Press.

Sturm, L. P., Windsor, J. A., Cosman, P. H., Cregan, P., Hewett, P. J., & Maddern, G. J. (2008). A systematic review of skills transfer after surgical simulation training. *Annals of Surgery*, *248*(2), 166-79.

Sudweeks, R. R., Reeve, S., & Bradshaw, W. S. (2005). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing*, *9*, 239-261.

Till, H., Myford, C., & Dowell, J. (2013). Improving student selection using multiple mini-interviews with multifaceted Rasch modelling. *Academic Medicine*, *88*(2), 216-223. doi: 10.1097/ACM.0b013e31827c0c5d

van der Vleuten, C. P. M., & Dannefer, E. F. (2012). Towards a systems approach to assessment. *Medical Teacher*, *34*(3), 185-186. doi: 10.3109/0142159X.2012.652240

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

Wilkinson, J. R., Crossley, J. G. M., Wragg, A., Mills, P., Cowan, G., & Wade, W. (2008). Implementing workplace-based assessments across the medical specialties in the United Kingdom. *Medical Education*, *42*(4), 364-373. doi: 10.1111/j.1365-2923.2008.03010.x

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*, 370. Retrieved from http://www.rasch.org/rmt/rmt83b.htm

Wright, B.D., & Masters, G.N. (1982). Rating scale analysis: Rasch measurement. Chicago, IL: MESA Press.

Figure 1. Variable Map Showing Student Clinical Ability Measures Estimated from Ratings Received in a Final-Year Ward Simulation Exercise (FYWSE), Assessor Severity Measures, Exercise Difficulty Measures, Domain Difficulty Measures, and Rating Scale Category Thresholds, all Reported on a Common Equal-interval Logit Scale.

```
+------------------------------------------------------------------------+
|Measr|+Students|-Assessors                  |-Exercises|-Domains  |Scale|
|-----+---------+----------------------------+----------+----------+-----|
|   6 +         +                            +          +          + (5) |
|     |         |                            |          |          |     |
|     |         |                            |          |          |     |
|   5 + .       +                            +          +          +     |
|     | .       |                            |          |          |     |
|     | .       |                            |          |          |     |
|   4 +         +                            +          +          + --- |
|     | .       |                            |          |          |     |
|     | *       |                            |          |          |     |
|     | *       |                            |          |          |     |
|   3 + *.      +                            +          +          +     |
|     | *.      |                            |          |          |     |
|     | *.      |                            |          |          |     |
|     | **      |                            |          |          |  4  |
|   2 + ***     +                            +          +          +     |
|     | *.      |                            |          |          |     |
|     | **      |                            |          |          |     |
|     | *.      |                            |          |          |     |
|   1 + **      + 5                          +          +          + --- |
|     | ***     | 3                          |          |          |     |
|     | *****   | 1  16 21 24 29             |          |          |     |
|     | ****.   | 2  7  8  10 11 13 14 19 26 | 1 2      |          |     |
*   0 * ******  * 18 22 23 31 32             *          * 10       *     *
|     | ****.   | 4  6  9  12 17 28 30       |          | 3        |  3  |
|     | ***.    | 27 33                      | 3        | 2  4  9  |     |
|     | *****.  | 15                         |          | 1  5     |     |
|  -1 + *****.  +                            +          + 6        +     |
|     | ***.    |                            |          |          | --- |
|     | **      |                            |          | 7  8     |     |
|     | **      |                            |          | 11       |     |
|  -2 + ***.    + 25                         +          +          +     |
|     | *.      |                            |          |          |     |
|     | **      |                            |          |          |  2  |
|     | *       |                            |          |          |     |
|  -3 + *       +                            +          +          +     |
|     | *       |                            |          |          |     |
|     | .       |                            |          |          | --- |
|     | .       |                            |          |          |     |
|  -4 +         +                            +          +          +     |
|     | .       |                            |          |          |     |
|     |         |                            |          |          |     |
|     |         |                            |          |          |     |
|  -5 +         +                            +          +          + (1) |
|-----+---------+----------------------------+----------+----------+-----|
|Measr| * = 2   |-Assessors                  |-Exercises|-Domains  |Scale|
+------------------------------------------------------------------------+
```

Table 1. Research Questions and Associated Propositions for Building a Validity Argument

Proposition	Research Questions
Proposition 1: The FYWSE measures critical behaviours that are required to provide high quality patient care in the context of the NHS.	**Question 1:** Were the simulation exercises used in the FYWSE developed to represent patient care reflective of the workplace in which the junior doctors will have to function?
Proposition 2: Students who perform better on the FYWSE will be more able to provide high quality patient care than those who do not perform well.	**Question 2:** Could the clinical performances observed and rated be considered a representative sample of the performances required in the clinical workplace? **Question 3:** Could the FYWSE identify those students who are competent in the clinical setting? To what extent has the instrument succeeded in separating students according to their levels of clinical performance?
Proposition 3: Students responded to the exercises in the manner that the assessment designers intended.	**Question 4:** Were there any students whose profiles of ratings showed an unacceptable amount of variability, suggesting that one (or more) assessors may have assigned rating(s) that were surprising or unexpected, given the other ratings that a student received?

Proposition 4: Assessors who judged the students were able to evaluate the students' clinical performances appropriately and fairly.	**Question 5:** Did the assessors have adequate background and clinical experience to carry out the clinical performance assessment? **Question 6:** Did the assessors differ in the severity with which they rated students' clinical performance? If they did differ in severity, how did those differences affect student scores? **Question 7:** Did the assessors use the categories on the rating scale in a consistent fashion?
Proposition 5: The instrument used to assess the students' clinical performance during the FYWSE functioned as intended.	**Question 8:** Are the three ward simulation exercises equal in difficulty, or are some of the exercises easier to get high ratings on than other exercises? **Question 9:** Are all domains included in the FYWSE equally discriminating? Which domains are harder to get high ratings on, and which are easier? **Question 10:** Did the 5-point scale used in the FYWSE function as intended?
Proposition 6: Eleven domains and two	**Question 11:** Are the scores obtained in the

assessors are sufficient to produce dependable, reliable/precise scores across replications of this assessment procedure.	FYWSE obtained from 11 domains and two assessors adequately reliable to make decisions about the students' clinical performance? **Question 12:** Which strategy would be more effective in improving the reliability of the FYWSE scores – increasing the number of assessors, increasing the number of domains, or some combination of the two?

Table 2. Parameter-level Mean-square Fit Statistics Showing the Number of Misfitting and Overfitting Students and Assessors, i.e. with Infit and Outfit Mean-square Values Less than 0.5 or Equal to or Greater than 1.5.

Mean-square Fit Statistics	Number of Infit MnSq Values (%)	Number of Outfit MnSq Values (%)
	STUDENTS	
< 0.5	16 (10.4%)	16 (10.4%)
0.5 – 1.49	114 (74.0%)	115 (74.7%)
1.5 - 1.99	15 (9.7%)	14 (9.1%)
> 1.99	9 (5.8%)	9 (5.8%)
	ASSESSORS	
< 0.5	0	1 (3.1%)
0.5 – 1.49	29 (90.6%)	28 (87.5%)
1.5 - 1.99	3 (9.4%)	3 (9.4%)
> 1.99	0	0

Table 3. Statistics by Rating Category

Category Label	Number of Times (%) a Category Was Used	Average Measure	Rasch-Andrich Threshold	Outfit Mean Square
1. Very poor	94 (3)	-2.10		1.2
2. Poor	463 (14)	-1.17	-3.28	1.0
3. Average	1135 (34)	.21	-1.38	.9
4. Good	1254 (38)	1.67	0.81	1.0
5. Outstanding	381 (11)	3.78	3.85	1.0

Table 4. Results of Generalizability Theory G-study ANOVA Showing Estimates of Variance Components (Assessor Nested in Student)

Source of variation	Degrees of freedom	Mean squares	Variance component	Percent of variability
e	2	28.31	0.02	1.68
s : e	151	9.45	0.38	38.11
a : s : e	154	0.72	0.03	3.33
d	10	18.31	0.06	5.74
ed	20	0.56	-0.00	0
sd : e	1510	0.67	0.16	15.86
ad : s : e	1540	0.36	0.36	35.32

a = assessor; e = exercise; d = domain; s = student

Table 5. Results of Generalizability Theory G-study ANOVA Showing Estimates of Variance Components (Assessor Crossed with Student)

Source	Degrees of freedom	Mean squares	Variance component	Percent of variability
e	2	28.31	0.02	1.69
s : e	151	9.45	0.38	37.71
a	1	31.95	0.01	0.59
d	10	18.31	-0.01	0
ea	2	0.18	-0.00	0
ed	20	0.56	-0.00	0
sa : e	151	0.52	0.03	2.76
sd : e	1510	0.67	0.23	22.51
sd	10	21.81	0.14	13.70
ead	20	0.28	0.00	0.12
sad : e	1510	0.21	0.21	20.93

a = assessor; e = exercise; d = domain; s = student

Table 6. Results of D-study (Assessor Nested in Student) Showing Projected Reliability with Varying Numbers of Assessors and Domains

Number of Domains	Number of Assessors	Generalizability Coefficient	Number of Assessors	Generalizability Coefficient
13	1	.80	2	.86
12	1	.80	2	.85
11	1	.80	2	.85
10	1	.78	2	.84
9	1	.77	2	.83
8	1	.76	2	.82
7	1	.74	2	.81
6	1	.72	2	.79