Adv in Health Sci Educ
DOI 10.1007/s10459-017-9771-4

Why assessment in medical education needs a solid foundation in modern test theory

Stefan K. Schauber[1] · **Martin Hecht**[2] · **Zineb M. Nouns**[3]

Received: 3 August 2015 / Accepted: 9 March 2017

Abstract Despite the frequent use of state-of-the-art psychometric models in the field of medical education, there is a growing body of literature that questions their usefulness in the assessment of medical competence. Essentially, a number of authors raised doubt about the appropriateness of psychometric models as a guiding framework to secure and refine current approaches to the assessment of medical competence. In addition, an intriguing phenomenon known as case specificity is specific to the controversy on the use of psychometric models for the assessment of medical competence. Broadly speaking, case specificity is the finding of instability of performances across clinical cases, tasks, or problems. As stability of performances is, generally speaking, a central assumption in psychometric models, case specificity may limit their applicability. This has probably fueled critiques of the field of psychometrics with a substantial amount of potential empirical evidence. This article aimed to explain the fundamental ideas employed in psychometric theory, and how they might be problematic in the context of assessing medical competence. We further aimed to show why and how some critiques do not hold for the field of psychometrics as a whole, but rather only for specific psychometric approaches. Hence, we highlight approaches that, from our perspective, seem to offer promising possibilities when applied in the assessment of medical competence. In conclusion, we advocate for a more differentiated view on psychometric models and their usage.

Keywords Measurement · Error · Assessment · Medical competence · Post-psychometric era · Case specificity · Latent variables

✉ Stefan K. Schauber
stefan.schauber@cemo.uio.no

[1] Centre for Educational Measurement at the University of Oslo (CEMO) and Centre for Health Sciences Education, University of Oslo, Oslo, Norway

[2] Department of Psychology, Humboldt–Universität zu Berlin, Berlin, Germany

[3] Institute of Medical Education, Faculty of Medicine, University of Bern, Konsumstrasse 13, 3010 Bern, Switzerland

Published online: 16 March 2017

Introduction

Psychometric models—developed within Classical Test Theory (CTT), Generalizability Theory (G Theory), and Item Response Theory (IRT)—have been widely employed in the field of research and assessment in medical education. They are routinely applied in high-stakes testing, such as the United States Medical Licensing Examination and the Medical Council of Canada Qualifying Examinations, as a means to secure and enhance trustworthiness and defensibility of assessments in medical education. Modern test theory also provides the basis for securing the defensibility of measurements in large-scale educational assessments, such as the Programme for International Student Assessment or the National Assessment of Educational Progress (Ray and Wu 2003; von Davier et al. 2006; Rutkowski et al. 2013). Decisions based on the results of these assessments can have far-reaching consequences, sometimes affecting a whole social system. For instance, vast educational reforms have been enacted as a consequence of students' inferior performances on the Programme for International Student Assessment (Grek 2009). In medical licensing examinations, assessments form not only the basis of decisions on an individual's career (becoming a doctor or not), but also are an integral part of securing the quality of the whole health care system (who becomes a doctor and who does not) (Norcini et al. 2011).

Despite the frequent use of current state-of-the-art psychometric models in the field of medical education, there is a growing body of literature that questions their usefulness in the assessment of medical competence. For instance, Schuwirth and van der Vleuten (2006) articulated a "plea for new psychometric models", and Hodges repeatedly advocated for the idea of a "post-psychometric era" (Hodges 2013; Eva and Hodges 2012). Essentially, these positions question the appropriateness of psychometric models as a guiding framework to secure and refine current approaches to the assessment of medical competence. The concerns raised by these authors are related to the reductionist approach that psychometric models take: the act of aggregating, summing, and thus reducing a rich variety of human behavior to a single number. However, this critical stance towards psychometrics is not limited to the field of medical education. Similar concerns have also been raised in the field of educational assessment. Indeed, since the 1970s, several authors have argued repeatedly against the meaninglessness of content-aptitude tests (McClelland 1973), the inappropriateness of norm-referenced testing in educational settings (Popham and Husek 1969), and the unquestioned elimination of items from tests based on statistical criteria alone (Goldstein 1979, 2012).

The criticism of the role of psychometrics in the assessment of medical competence has been accompanied by a redefined approach to assessment in general. Schuwirth and van der Vleuten (2011) developed the concept of 'programmatic assessment', in which students' learning and professional development plays a crucial role. Hodges (2013) put forward the idea of assessment as a gestalt: a meaningful whole that is "…more than its parts." Although these authors are clearly critical of psychometrics, they do not reject its use *per se*; but it plays a minor role in their conception of assessment. In a recent publication, van der Vleuten et al. (2014) suggested that high-stakes decisions could legitimately be based on professional judgment, and that the trustworthiness of such decisions should be ensured by expert committees. Judgment in this context does not translate to intuitive or ad-hoc judgments, but rather to elaborate, deliberate professional evaluations and decisions, comparable to ethical review boards. In the perspectives advocated by Schuwirth and van der Vleuten (2011), Hodges (2013), and Cook et al. (2016), professional judgment has a central function that used to be reserved for psychometric methods: the role

of ensuring the trustworthiness and credibility of assessments and accompanying high-stakes decisions. Indeed, the trust in psychometric methods to accomplish this task seems to have weakened substantially. In 2014, van der Vleuten asked, "[s]hould we replace psychometric theories with an interpretative theory" (p. 235), a theory rooted in qualitative research methods (Driessen et al. 2005), and possibly put an end to "…the dominance of the psychometric discourse" (Hodges 2013)?

From a psychometric perspective, this critical stance against the field, which is devoted to the development of approaches that ensure fair and defensible assessment, may be rather irritating. However, in a broader understanding, psychometric methods are only one part of a much larger framework that allows us to make inferences about students' competence in a systematic, scientific manner. Assessment, from the perspective of modern test theory, is a systematic approach that encompasses everything from defining the purpose of an assessment, to specifying and developing test content, to reaching conclusions such as pass and fail decisions (Wilson 2005). Obviously, assessment as a whole cannot be carried out by the application of a statistical formula alone, but rather by the process of rigorous test development, in which the actual tests or exams administered are just part of the story. Indeed, discussions in medical education so far have only focused on a specific part of the general measurement process, that is, on the statistical combination of information. Although this is only a small part of an overarching process, the way in which information is combined, weighted, and summarized is crucial, as such procedures ultimately affect decisions that concern individuals. In that respect, psychometrics can be regarded as a set of rules for combining data in a way as simple, or as complex as may be deemed necessary to reach consistent and reproducible inferences on individuals. Ultimately, psychometric models can be used to make consequential decisions on individuals fully traceable, debatable, and revisable.

The traceability of any single decision down to the individual responses in a particular exam is a key advantage of an approach firmly rooted in modern test theory. A lack of transparency or insight is, at the same time, of grave concern regarding the use of human, and even expert, judgment in high-stakes assessment. Research on human judgment in various domains has repeatedly shown that individuals are hardly aware of how they weigh and combine available information to form a judgment or decision—decision making is not fully conscious, and therefore it is potentially error-prone. As stated by Evans et al. (2003), "If experts lack self-insight into the processes underlying these judgments, they may be unconsciously biased." (p. 608) Experts do make fairly accurate judgments and decisions every day; however, previous research has indicated that such accuracy may depend substantially on both the structure of the situation and the object of judgment (Hammond et al. 1987). For instance, a recent study found that experts may be rather accurate in recognizing close-to-ideal performances but are less able to sufficiently discriminate between low-level performers (Larson and Billeter 2016). In addition, the accuracy of an expert's judgment is usually outperformed by mathematic models based on experts' implicit rules (Goldberg 1970; Karelaia and Hogarth 2008). In this same vein, a meta-analysis by Karelaia and Hogarth (2008) stated that, "…decision making procedures […] should be replaced by models derived from human decision makers." (p. 407). In conclusion, while expert judgment is crucial, research highlights that the accuracy and consistency of actual decisions remains an issue (Slovic and Lichtenstein 1971; Kaufmann and Athanasou 2009), but it can be enhanced by expert-informed and systematically applied mathematical rules.

Indeed, one way to conceive of psychometrics is as a set of systematically applied mathematical rules, and the corresponding methods to investigate the appropriateness of

those rules. As noted above, one critique of quantitative methods in particular has been that these 'rules', the process of assigning numbers to observed performance, are reductionist in nature and that, consequently, this approach would assume that it is sensible to aggregate or summarize across observations (Hodges 2013). In this regard, there is one issue that is specific to the controversy on the use of psychometric models for the assessment of medical competence: a finding usually referred to as 'case specificity'. Broadly speaking, case specificity is the finding of instability of performances across clinical cases, tasks, or problems. Findings of case specificity are usually reported in two types of studies. First, as noted by Norman (2008), results from correlational studies indicate that associations between performances are often low (see also Elstein 1978; Norman et al. 1985; Roberts and Norman 1990). Second, psychometric studies have repeatedly indicated that the amount of unexplained variance in various assessment scenarios is comparably large. Indeed, 60–70% of the total variance often remains unexplained (Brannick et al. 2011; Wrigley et al. 2012; Ricketts et al. 2010; Dory et al. 2010; Norman et al. 2006; Colliver et al. 1990; Jarjoura et al. 2004; De Champlain et al. 1999; Swanson et al. 1995; Richter Lagha et al. 2012). Taken together, the finding of case specificity is synonymous with a low degree of within-person consistency of performances across items or cases within and across assessments. Similar results have also been found in other domains (Shavelson et al. 1993, 1999). Such a low degree of consistency is associated with a struggle to arrive at defensible assessments. "That certain performance scores do not generalize indicates that the measurement process has low reliability and validity, and may highlight the need to develop alternative measurement methods that perform better." (Kreiter 2008) But does this also warrant the conclusion that psychometric reductionism isn't appropriate for the assessment of medical competence?

Indeed, the finding of case specificity may play a crucial role in the reservations psychometrics is faced with. Case specificity has been called the "…one truth in medical education" (John Norcini in Eva 2011, p. 22), since the associated variability of performances is found almost everywhere and across many contexts (Eva 2003). Knowing this, the remark that "…a scientific model capable of explaining only such a small portion of the observed variance is at best a moderately strong model" (Schuwirth 2009, p. 299) is highly interesting. Schuwirth and van der Vleuten (2006) also argued that core concepts of psychometric theory, such as latent variables, might not be meaningful in the assessment of medical competence, stating "we […] think the assumption that they [i.e., the aspects of medical competence] can be treated as latent constructs is incorrect…" (p. 297), amongst other reasons, because "…[i]n this model, constructs are used as generic, stable and homogenous characteristics" (p. 296). Clearly, the finding of case specificity seems to represent quite the contrast to the notion of stability. Taken together, this raises the question of whether the finding of case specificity can be regarded as an empirical argument for the inadequacy of psychometric models in the assessment of medical competence.

The purpose of this article is to argue that modern test theory is imperative for fair and defensible assessment. The definition of modern test theory we use is that of as an overarching framework of assessment ranging from theoretic considerations on the phenomena of interest to psychometric modelling and, ultimately, actual inferences from test scores (Wilson 2005). Specifically, we argue that the finding of case specificity cannot be regarded as empirical evidence to support the inappropriateness of the general statistical methods applied in assessments in medical education. However, we note that the two concepts that seem to be challenged by the persistent finding of case specificity are the framework of latent variables and the concept of measurement precision (and particularly the methods used to estimate the reproducibility of test results as formulated in both CTT

and G Theory). Consequently, this article is structured as follows: first, we will delineate the concept of latent variables and measurement precision; second, we will try to highlight the extent to which the finding of case specificity represents a challenge in psychometrics. Finally, we will conclude in a discussion and argue that assessment in medical education needs a solid foundation in modern test theory.

Latent variables

Presumably without even knowing, we deal with latent variables on a daily basis. For instance, if we observe a particular student cheating repeatedly in exams and suspect that he/she plagiarized a homework assignment, it is likely that we will come to the conclusion that this student is in fact a 'cheating' person. In short, we tend to attribute the observed behavior (e.g., the event of bringing a crib sheet to an exam) to a trait, that is, a stable facet of a person's character or a general disposition to act in some characteristic way across situations (e.g., being a 'cheater'). In such a case, we would not really be surprised if this particular student is caught cheating again. Indeed, we might have a more or less explicit expectation and would keep a close eye on that student during the next exam. However, we cannot see, feel, taste, or hear this trait; it is to some extent hidden or latent. By attributing a person's behavior to their character, we implicitly assume a causal force, i.e., that this trait is the reason why she/he brought a crib sheet to the exam, plagiarized, or even used 'made-up' laboratory data. Whenever observed behavior leads to inferences about how a person may act at future occasions, those inferences will be of a latent trait or variable. In this respect, the concept of latent variables is not a uniquely psychometric one.

The most worthwhile feature of latent variables is often seen in "…the intuitive appeal of explaining a wide range of behaviors by invoking a limited number of latent variables." (Borsboom et al. 2003, p. 203). The psychometric literature describes the model underlying such a rationale (i.e., observations caused by a latent variable) as a reflective measurement model (Edwards and Bagozzi 2000; Borsboom et al. 2003; Bollen and Lennox 1991). Theoretically, the observed behaviors (performances on items, ratings on check-lists, responses to questionnaires, etc.) are conceived of as indicators of the effect of the latent variable. Similar inferences may be made in jury trials when the legal proceedings rely on circumstantial evidence. This may result in convictions such as: "Given his dishonesty in other cases, we regard it as highly likely that he also misappropriated research grants". The behavior at one point in time is used as a predictor of a behavior or response at another point, since they all are evoked by the same latent variable. This implicit causal attribution (e.g., dishonesty causing cheating, betrayal, plagiarizing) is also depicted in the graphical notation system of structural equation models, in which arrows point from a latent variable (depicted by ellipses) to manifest indicators (depicted by squares), as shown in Fig. 1. The reflective measurement model implies that items that are combined together—observations of the same class—share some meaningful or useful redundancy (Edwards 2011; Bollen and Lennox 1991) with regard to the latent trait. Which observations constitute a class or entity and thus share meaningful redundancy is first and foremost a theoretical question, a decision made by the researcher. However, that decision has to be justified empirically.

One way to justify a theoretical model in a latent variable analysis is to compare the specified model (and alternative models) to the observed data. The method used to explore the alignment between a model and the data is a central concept in statistical modelling and is usually referred to as the evaluation of model fit. According to Sijtsma (2006), psychometric models can be conceived of as a mathematical condensation of the observed data: "latent variables—latent traits, factors, and latent classes—are summaries of the data

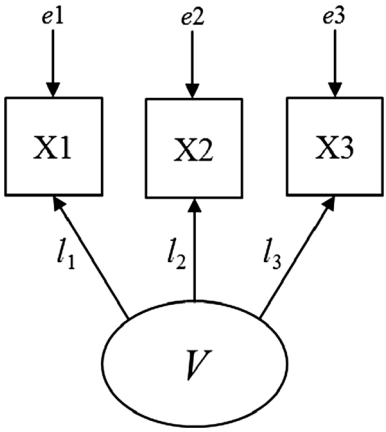

Fig. 1 A reflective measurement model. V signifies the latent variable; l1–l3 the respective factor loadings; X1–X3 are the manifest (observed) variables; e1–e3 is measurement error

and nothing more."(p. 452) A model is always a simplification; hence, deviations from that formalization are expected. Indeed, a model that would account for all details or specifics in the data will hardly be generalizable to other occasions or instances. Nevertheless, more complex models are at least equally good or better in explaining the data, as they account for more idiosyncrasies in the observations, which would otherwise remain unexplained or un-modelled. One implication of the approach to evaluate model fit is that, for a set of models applied to an identical dataset, each model may explain the data to varying degrees. Investigating the relative fit of rivaling models may be one approach to justify the way assessment data is summarized and consequently how corresponding high-stakes decisions are made.

The crucial part in assessing model fit is to compare the relative efficiency of different, theoretically sound models to explain the observed data. When these comparisons are performed, the model ultimately chosen can be regarded as the most appropriate summary of the observed information. For instance, in typical multiple-choice exams in undergraduate medical education, items will vary in their difficulty, that is, the relative frequency of correct answers. In addition, the usefulness of different items to discriminate between students varies, that is, getting particular items right or not may be differently predictive of students' overall exam performance: Some items reveal more about who is a high-performing and who is a low-performing student. In the context of IRT, there is an explicit choice to account for these two distinct item properties, and inferences on students' ability might be based on the ability estimates derived from the better fitting IRT model. In this scenario, at least two models might be specified and applied to the data. First, a 1-parameter-logistic model, which accounts for differences in item difficulties while assuming item discriminations to be constant. Second, a 2-parameter-logistic model, which estimates both difficulty and discrimination parameters for the items in the exam. The explanatory power of both models can then be compared using various model fit criteria. The better fitting model may then be regarded as a more accurate summary of the available data. Thus, an objective criterion helps to determine—and justify—how the patterns of students' responses can be most adequately aggregated.

However, determining which model is most suitable is not only a function of statistical criteria, but also of the purpose of an assessment. Therefore, a simpler and possibly less well fitting model may be chosen over a more complex model. For instance, one consequence of accounting for item discrimination in an IRT model is that responses are

weighted differently. Thus the estimate of students' abilities derived from individual response patterns is not only a function of how many items, but also of which items students answered correctly. In some research scenarios, the weighting of items might be of particular interest. Still, in a high-stakes exam it might be problematic to explain why two students got different grades although they solved the same number of items correctly. In addition, legal requirements might determine the model applied, particularly in contexts where local regulations define how to arrive at a test score. If number correct scores have to be used, a 1-parameter-logistic model is required, regardless of whether a more complex model would be statistically more adequate. The consequences of decisions on how to combine or weigh items (by the researcher or regulation authorities) can and should be investigated. It's not only the model that can be put to the test but also the actual consequences for individual test takers.

To summarize, a central concept in latent variable analyses is that observations are combined to form a latent variable. The advantage of this approach is that it gives the ability to explain otherwise unrelated observations, predict behavior, or, for that matter, performances. In order to evaluate the degree of suitability of a particular model the fit of the model to the data must be examined. The choice of a particular model can be as much a matter of professional judgment as one of statistical comparisons or the context of application. If, however, a less well fitting model is selected—especially when it is used for practical purposes such as routine assessments—the benefits, drawbacks, and implications of possibly incorrect inferences can be analyzed and evaluated. Latent variable models—summaries of the data—can be tested based on both their statistical properties and their practical consequences.

Measurement precision

Generally speaking, the topic of measurement precision encompasses all efforts that aim to estimate the reproducibility of measurements (e.g., exam results, ratings, classifications), that is, an estimate of the extent to which "…results of the assessment would be the same if repeated under similar circumstances." (Norcini et al. 2011). From a psychometric perspective, such an estimate is important in securing the trustworthiness of claims that are based on test scores (e.g., inferences of proficiency) (Kane 1996, 2013; Messick 1989). One approach to determine the reproducibility of test results has been developed in G Theory (Brennan 2001). Briefly, G Theory "…pinpoints the sources of measurement error, disentangles them, and estimates each one." (Webb et al. 2006). Measurement error and measurement precision are closely related concepts, since examining the sources of measurement error may help to build more replicable assessments. In this regard, G Theory is often considered to be more flexible than CTT (Crossley et al. 2002). This, because G Theory conceives of measurement error as multi-faceted, which is in contrast to the single general error term in CTT "observed score = true score + error" formulation. G Theory aims to 'unpack' (Zumbo 2006) this single error term further.

Although a comprehensive overview of G Theory is beyond the scope of this paper, the concept of multi-faceted measurement error is of particular importance in the current context. In G Theory, measurement error is related to inconsistencies, or a lack of reproducibility, in the data. In most assessments, various facets or factors contribute to variability in tests scores. In a simple multiple-choice exam, where students answer a number of unique items, variability in scores derives from differences in students' ability—students differ from each other with respect to their overall test score. This between-person variability constitutes the construct of interest whenever the purpose of the assessment is to

differentiate between high-performing and probably lower-performing students. Furthermore, individual test items typically have different levels of difficulty (i.e., they vary with respect to the relative number of students that answered them correctly). Hence, between-item variability is a second source of variance in the assessment procedure. Additionally, as in any other statistical model, there is the assumption of an unexplained remainder, that is, residual variance. Residual variance "…represents what is commonly thought of as error of measurement, combining the variability of performance to be expected when an individual can sometimes exceed his norm by gaining insight into a question and sometimes fall short because of confusion, a lapse of attention, and so forth." (Cronbach and Shavelson 2004) What usually cannot be investigated in a typical exam is the interaction between the student and the item, meaning we don't know if students would receive the same score if they had to answer that same item again. In a typical exam, this interaction cannot be disentangled from the residual component; thus student-item-interaction and residual variance combined form a third source of variability. In summary, in typical applications, the between-student differences (the student facet) are the source of variation of interest, while the residual component is always regarded as measurement error and represents the unexplained variability in the data. However, which variance component represents error variance and which forms the construct of interest is a matter of theoretical consideration.

Whether or not a specific facet constitutes measurement error is a substantive decision and is specific to the context of application, which can be illustrated by two examples. In the first example, a paper-and-pencil exam is administered using different booklets that contain different overlapping item subsets. Usually, the booklets should not be related to differences in performance between students. However, if an analysis found a substantial variance attributable to the booklets, this facet may interfere with the measurement of the construct of interest (student performance) if not adequately disentangled (e.g., Hecht et al. 2015). In the second example, an exam includes different subjects (e.g., anatomy, physiology, biochemistry), and students may have different levels of proficiency in these subjects. If so, a substantial student-subject interaction would be estimated. In this context, performances may not be consistent across the exam as a whole, but the variability across subjects or domains may still be regarded as the construct of interest. Hence, inconsistencies would not generally be considered error variance, but whether they are or not is a matter of both theoretical consideration and of the inferences to be made from the assessment. In this respect, and similar to latent variable modelling, G Theory allows the researcher to formulate and evaluate expectations on which observations are deemed to be observations from the same class and estimate the degree to which observations within that class are replicable.

Both G Theory and latent variable modelling represent statistical techniques that aim to explain systematic variation, that is, to find some sort of consistency and replicability in the data. Consequently, it may not be surprising that G Theory can be regarded as a special type of latent variable model (Skrondal and Rabe-Hesketh 2007; Zumbo 2006; Marcoulides 1996). Consider a clinical encounter where a physician sees a patient with diabetes. In such an encounter, the physician may first need to give information on how to handle diabetes in everyday life, and then prescribe the correct medication, drawing on knowledge of the mechanisms of insulin release to do so. An assessment that covers these three domains (patient education, medication, and mechanisms) administered to a group of physicians might indicate different patterns of performance; for example recalling the influence of sulfonylurea on insulin release may have become challenging for some physicians. For others, after having talked to hundreds of patients, explaining the necessary

change in diet may have become rather easy. Table 1 illustrates these different patterns of performance: different physicians perform differently on each task—but consistently within tasks. Inconsistency across all observations may then be systematic and reflected in several subdomains of competence. In such a scenario, a variant of a G Theory model could indeed be fit to the data using a latent variable framework, as shown in Fig. 2 (Marcoulides 1996). Complete translation between these approaches is not possible, but there is substantial theoretical and analytical overlap (Webb et al. 2006).

The finding of case specificity and its relation to criticisms of psychometrics

While the concept of replicability or stability across a specific set of observations plays a crucial role in the psychometric concepts discussed, the recurrent finding of case specificity suggests that the opposite is the "one truth" (John Norcini in Eva 2011, p. 22) in medical education, i.e., performances observed in assessments of medical competence do not seem stable or reproducible across contexts, tasks, cases, and so forth. As noted in the introduction, this finding might play a crucial role in the criticisms levelled at psychometric models. The widespread finding of case specificity might be regarded as empirical evidence of 'misfit' of psychometric models and may therefore underpin the argument that "…a scientific model capable of explaining only such a small portion of the observed variance is at best a moderately strong model." (Schuwirth 2009, p. 299).

While the finding of case specificity remains an interesting phenomenon, psychometrics and latent variable modelling offer various approaches to acknowledge the phenomenon of comparably large amounts of unexplained variance by exploring various considerations. First, the phenomenon of case specificity might not be the result of error-prone assessment procedures but rather of a multitude of unmeasured factors that could be taken into account and investigated with psychometric models (Colliver et al. 1990; Kreiter and Bergus 2007; Crossley 2010). Second, additional variance may be explained by assuming a more complex structure of medical competence, which may be understood as a multi-dimensional construct (Wimmers and Fung 2008; Wimmers et al. 2007; Mattick et al. 2008). At the very least, results from performance-based assessments may be best modelled using more complex analysis techniques (Keller et al. 2010). However, most psychometrically driven efforts to explain the finding of case specificity have gotten stuck between those two

Table 1 Example for assessment results that show variability across domains and consistency within domains

Person	Diabetes								
	Patient education			Medication			Mechanism of insulin release		
	1	1′	1″	2	2′	2″	3	3′	3″
1	1	1	1	0	0	0	0	0	0
2	0	0	0	1	1	1	1	1	1
3	0	0	0	0	0	0	1	1	1
4	1	1	1	1	1	1	0	0	0
5	0	0	0	1	1	1	1	1	1
6	1	1	1	0	0	0	0	0	0

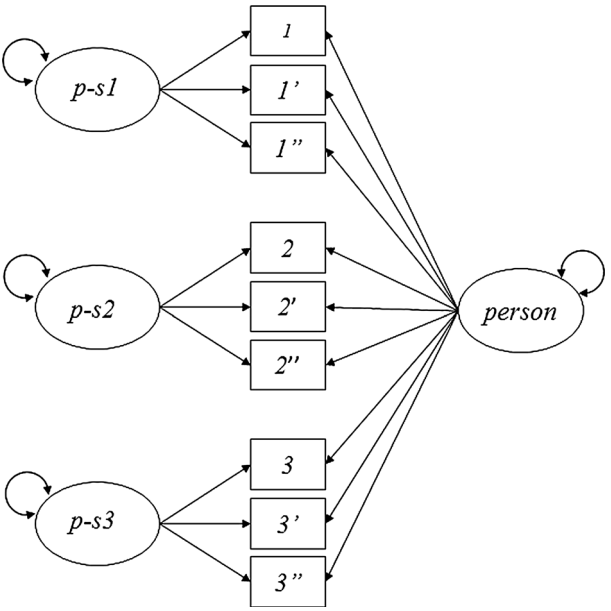

Fig. 2 G Theory model from a latent variable perspective for data in Table 1, based on (Marcoulides 1996). *Directed arrows* are fixed to one. Variances of the skill-specific latent variables (p-s1-3) are constrained to be equal. Then, the variances of the latent variables—the *double headed arrows* in the graph—equal the variance components in a generalizability study which would include person, person-skill-interaction and residual variance components (Marcoulides 1996)

positions. Although—sometimes slight—variations in context have been shown to play a critical role in determining performances both in cognitive psychology research (Godden and Baddeley 1975; Leight and Ellis 1981; Goodwin et al. 1969; Kotovsky et al. 1985; Gick and Holyoak 1980) and in medical education (Durning et al. 2012), a univocal consensus on the actual causes of case specificity seems to be lacking. As noted in the introduction, case specificity is frequently assigned to the finding that up to 70% of the total variance in a particular assessment context remains unexplained. While such a share may sound irritating, a critical question that has rarely been addressed is: how small could the proportion of residual variance theoretically be? Put differently, is 70% really large?

The expectance of consistency and explained variation are as dependent on the theoretical model as the expectance of residual variation. In deterministic models, such as CTT or G Theory models, the proportion of residual variation in the total variance could be virtually 0%, and ideally this would be reached. Table 2 illustrates such a response pattern, where scores across observations are perfectly stable. In such a model, any observed score

Table 2 Ideal data pattern according to a deterministic response process - 0% residual variance

Person	Items						Average score
	1	2	3	4	5	6	
1	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0
4	1	1	1	1	1	1	1
5	1	1	1	1	1	1	1
6	1	1	1	1	1	1	1
Average score	.5	.5	.5	.5	.5	.5	.5

Residual variance = <.001, item variance = <.001, between-person variance = .043

carries all information about the other observations. In contrast, in probabilistic models such as IRT models, the relation between a latent variable and the observed outcome (e.g., diagnosing a case correctly or not) is formulated in terms of chances for success. If such a probabilistic process is assumed, data can be in perfect agreement with the model, but the amount of residual variance, from the perspective of a deterministic model will be comparably high. To substantiate this claim, we simulated data from a model according to a uni-dimensional and probabilistic response process, such as that found in a Rasch model.[1] We selected parameters that would mimic typical assessment scenarios in medical education. The results represent summary statistics for 5000 drawn samples, with 300 persons and 100 items in each sample. Across the 5000 samples, percent-correct scores ranged, on average, between 48 and 92%. The grand mean of the average within-sample test-score was 74% correct. A G Theory-based analysis, conducted in the R language for statistical computing (R Core Team 2013; Bates et al. 2015), showed that on average, 2.5% of the total variance was attributable to persons, 25% to items, and 72.5% to residual variance. This pattern of variance components would likely be interpreted as the finding of case specificity. However, from the perspective of a probabilistic model, 70% residual variance is not large, but can be readily expected in the given scenario while ruling out the possibility of an effect of multi-dimensionality or the influence of unmeasured factors.

A relatively large share of residual variance may not pose a psychometric issue at all, but rather may point to an inherently stochastic relation between ability and success (or failure) on items, cases, or tasks. This perspective is generally in alignment with a long tradition of research on judgmental processes (Slovic and Lichtenstein 1971; Hammond et al. 1964; Cooksey 1996) and echoes a recent conception of the process of diagnostic inference. Hertwig et al. (2013) argued that "…[b]ecause cognition and perception are probabilistic and based on imperfect cues, there is a natural limit to how accurate they can be. Inevitable though errors may be, they do not reflect a failure of the inferential system but a probabilistic environment that is not perfectly predictable from the available cues." (p. 534) If such a probabilistic environment can legitimately be assumed, methods developed within IRT may be a theoretically adequate fit to model such inherently stochastic processes. The proposition of the "probabilistic nature of diagnostic inference" (Hertwig et al. 2013, p. 534) suggests that we may have become so used to explaining the finding of case specificity from a substantive point of view, as a result of the complexities in assessing medical competence, that it is very hard to recognize the possibility that those patterns could also be governed by a very simple, but probabilistic, response process.

The role of psychometrics in the assessment of medical competence

The considerations delineated in the sections above suggest that probabilistic models, as developed within IRT, could contribute to a psychometric model that fits the assessment of medical competence better than traditional, deterministic models in both a theoretical and an empirical sense. However, models created within the framework of IRT often face concerns regarding their applicability, especially in small-scale scenarios (e.g., 200 students or less). Knowing this, it may be worthwhile to draw on earlier investigations on the applicability of such psychometric approaches in small-sample scenarios, which point out that simple IRT models may be legitimately used in sample of about 100 students (Jones et al. 2006). Furthermore, for the specific context of assessment in medical education, simulation studies, and secondary analyses of actual examination data seem to be

[1] R scripts for this simulation are available upon request from the corresponding author.

promising analytic strategies to arrive at practical recommendations for the use of probabilistic measurement models in routine applications in medical schools.

We furthermore propose that several concepts developed in modern test theory can be of great benefit within the framework of programmatic assessment as delineated by Schuwirth and van der Vleuten (2011). We want to briefly highlight three specific concepts that align with the implications of programmatic assessment. First, the idea of tailoring the assessment to the individual student aligns well with the concept 'information' in a psychometric sense (Mellenbergh 1996). This allows us to determine at which point enough data is available on a person's ability to give reliable feedback or make defensible decisions. Second, statistical methods developed for clinical trials lean on the concept of sequential sampling and purposeful sampling (Bartroff et al. 2013); the idea of obtaining and sampling data where it is deemed necessary and most informative is also evident in the framework of programmatic assessment. Third, Bayesian approaches capitalize on the idea that prior information on performances is usually available and, again, offer a rule for combining data from different sources in a systematic manner. Prior information may stem from expert judgment but could then be applied and investigated systematically. Furthermore, Bayesian approaches are also an interesting alternative in situations where traditional estimation techniques are limited and may be especially useful in small-sample scenarios. A careful delineation of those approaches is beyond the scope of this paper, but these concepts may constitute the next steps to following Schuwirth and van der Vleuten's (2006) call for new psychometric models, and to exploring a "probabilistic or Bayesian approach" (p. 300).

Discussion

A decade after the call for new psychometric models by Schuwirth and van der Vleuten (2006), psychometricians seem to have only rarely responded to the problems of these models. In place of this, a body of literature has accumulated that has been skeptical about the role of psychometrics in the assessment of medical competence. The frequent occurrence of case specificity may also have fueled doubts on the applicability of psychometric models within the medical education community. Against this background, this article aimed to summarize, review, and illustrate the concepts that are frequently referred to in this discussion. We started with the description of the commonly reported inconsistencies in such measurements, often attributed to the finding of case specificity. Subsequently, we aimed to delineate the possible conflicts between the underpinnings of core concepts in psychometric theory on the one hand, and the persisting finding of case specificity on the other hand. We highlighted that expectancies of stability or variability are, first and foremost, a matter of theoretical consideration and inferences from assessments.

We want to stress that psychometric methods and modern test theory are in general much more flexible than usually described. However, we agree that a strict CTT approach might be less suitable for scenarios that are typically of interest in the assessment of medical competence. Importantly, psychometric theories seem to be in a phase of unification, which comes with an increasing breadth of possible analytic approaches. For instance, under the framework of generalized linear mixed models, methods that interweave the estimation of variance components (as in G Theory) and IRT have now become easily available and applicable (Doran et al. 2007). Hence, the stark contrast between psychometric traditions seems to have vanished. In the light of such a broadened toolbox,

the substantive question that needs to be addressed before fitting a model to the data and drawing inferences from it is related to the conceptual appropriateness of the measurement model itself: "Psychometric techniques and models have great potential for improving measurement practice […] but only if they are driven by a substantive theory of response processes."(Borsboom et al. 2004, p. 1070).

Conclusion

From our perspective, a solid foundation in modern test theory, encompassing theoretic considerations on the phenomena of interest as well as psychometric modelling, is indispensable in order to secure the trustworthiness and defensibleness of high-stakes decision making. However, the beneficial application of psychometric methods within the context of the assessment of medical competence is not without its challenges. These may include more technical topics, such as accessibility or applicability of certain procedures, as well as more conceptual considerations regarding the theoretical appropriateness of particular methods. To promote the beneficial application of modern test theory, it seems crucial to foster assessment literacy among lecturers, teachers, assessment staff, and researchers. This is not only true in the context of assessment in medical education (Popham 2009; Borsboom 2006). Importantly, quantitative techniques have some features that do not translate well into more qualitative or interpretative approaches. Amongst the most important are the possibility of fully transparent—and thus debatable—decision processes, and the opportunity to probe different models and consequences derived from these models. This is especially important as the final outcome, the final measure of competence, is inevitably a dichotomous, quantitative one. In the long run, a student will either pass or fail his or her studies as a whole—and there is no room for interpretation in between.

References

Bartroff, J., Lai, T. L., & Shih, M.-C. (2013). *Sequential experimentation in clinical trials*. New York, NY: Springer.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*. doi:10.18637/jss.v067.i01.

Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin, 110*(2), 305.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika, 71,* 425–440. doi:10.1007/s11336-006-1447-6.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110,* 203–219. doi:10.1037/0033-295X.110.2.203.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111,* 1061–1071. doi:10.1037/0033-295X.111.4.1061.

Brannick, M. T., Erol-Korkmaz, H. T., & Prewett, M. (2011). A systematic review of the reliability of objective structured clinical examination scores. *Medical Education, 45,* 1181–1189. doi:10.1111/j.1365-2923.2011.04075.x.

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

Colliver, J. A., Markwell, S. J., Vu, N. V., & Barrows, H. S. (1990). Case specificity of standardized-patient examinations: Consistency of performance on components of clinical competence within and between cases. *Evaluation & the Health Professions, 13,* 252–261. doi:10.1177/016327879001300208.

Cook, D. A., Kuper, A., Hatala, R., & Ginsburg, S. (2016). When assessment data are words: Validity evidence for qualitative educational assessments. *Academic Medicine*. doi:10.1097/ACM.0000000000001175.

Cooksey, R. W. (1996). The methodology of social judgement theory. *Thinking & Reasoning, 2,* 141–174. doi:10.1080/135467896394483.
Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement, 64,* 391–418. doi:10.1177/0013164404266386.
Crossley, J. G. M. (2010). Vive la difference! A recall from knowing to exploring. *Medical Education, 44,* 946–948. doi:10.1111/j.1365-2923.2010.03786.x.
Crossley, J., Davies, H., Humphris, G., & Jolly, B. (2002). Generalisability: A key to unlock professional assessment. *Medical Education, 36*(10), 972–978.
De Champlain, A., MacMillan, M. K., King, A. M., Klass, D. J., & Margolis, M. J. (1999). Assessing the impacts of intra-site and inter-site checklist recording discrepancies on the reliability of scores obtained in a nationally administered standardized patient examination. *Academic Medicine, 74*(10), S52–S54.
Doran, H., Bates, D., Bliese, P., & Dowling, M. (2007). Estimating the multilevel Rasch model: With the lme4 package. *Journal of Statistical Software*. doi:10.18637/jss.v020.i02.
Dory, V., Gagnon, R., & Charlin, B. (2010). Is case-specificity content-specificity? An analysis of data from extended-matching questions. *Advances in Health Science Education, 15,* 55–63. doi:10.1007/s10459-009-9169-z.
Driessen, E., van der Vleuten, C. P. M., Schuwirth, L., van Tartwijk, J., & Vermunt, J. (2005). The use of qualitative research criteria for portfolio assessment as an alternative to reliability evaluation: A case study. *Medical Education, 39,* 214–220. doi:10.1111/j.1365-2929.2004.02059.x.
Durning, S. J., Artino, A. R., Boulet, J. R., Dorrance, K., van der Vleuten, C. P. M., & Schuwirth, L. (2012). The impact of selected contextual factors on experts' clinical reasoning performance (does context impact clinical reasoning performance in experts?). *Advances in Health Science Education, 17,* 65–79. doi:10.1007/s10459-011-9294-3.
Edwards, J. R. (2011). The fallacy of formative measurement. *Organizational Research Methods, 14,* 370–388. doi:10.1177/1094428110378369.
Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods, 5*(2), 155–174.
Elstein, A. S. (1978). *Medical problem solving: An analysis of clinical reasoning*. Cambridge, MA: Harvard Univ. Press.
Eva, K. W. (2003). On the generality of specificity. *Medical Education, 37,* 587–588. doi:10.1046/j.1365-2923.2003.01563.x.
Eva, K. (2011). On the relationship between problem-solving skills and professional practice. In C. Kanes (Ed.), *Elaborating professionalism* (Vol. 5, pp. 17–34, Innovation and change in professional education). Dordrecht: Springer.
Eva, K. W., & Hodges, B. D. (2012). Scylla or Charybdis? Can we navigate between objectification and judgement in assessment? *Medical Education, 46,* 914–919. doi:10.1111/j.1365-2923.2012.04310.x.
Evans, J. S. B. T., Clibbens, J., Cattani, A., Harris, A., & Dennis, I. (2003). Explicit and implicit processes in multicue judgment. *Memory & Cognition, 31,* 608–618. doi:10.3758/BF03196101.
Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology, 12,* 306–355. doi:10.1016/0010-0285(80)90013-4.
Godden, D. R., & Baddeley, A. D. (1975). Context-dependent memory in two natural environments: On land and underwater. *British Journal of Psychology, 66,* 325–331. doi:10.1111/j.2044-8295.1975.tb01468.x.
Goldberg, L. R. (1970). Man versus model of man: A rationale, plus some evidence, for a method of improving on clinical inferences. *Psychological Bulletin, 73,* 422–432. doi:10.1037/h0029230.
Goldstein, H. (1979). Consequences of using the Rasch model for educational assessment. *British Educational Research Journal, 5,* 211–220. doi:10.2307/1501031.
Goldstein, H. (2012). Francis Galton, measurement, psychometrics and social progress. *Assessment in Education: Principles, Policy & Practice, 19*(2), 147–158.
Goodwin, D. W., Powell, B., Bremer, D., Hoine, H., & Stern, J. (1969). Alcohol and recall: State-dependent effects in man. *Science, 163,* 1358–1360. doi:10.1126/science.163.3873.1358.
Grek, S. (2009). Governing by numbers: The PISA 'effect' in Europe. *Journal of Education Policy, 24,* 23–37. doi:10.1080/02680930802412669.
Hammond, K. R., Hamm, R. M., Grassia, J., & Pearson, T. (1987). Direct comparison of the efficacy of intuitive and analytical cognition in expert judgment. *IEEE Transactions on Systems, Man, and Cybernetics, 17,* 753–770. doi:10.1109/TSMC.1987.6499282.
Hammond, K. R., Hursch, C. J., & Todd, F. J. (1964). Analyzing the components of clinical inference. *Psychological Review, 71,* 438–456. doi:10.1037/h0040736.
Hecht, M., Weirich, S., Siegle, T., & Frey, A. (2015). Modeling booklet effects for nonequivalent group designs in large-scale assessment. *Educational and Psychological Measurement, 75,* 568–584. doi:10.1177/0013164414554219.

Hertwig, R., Meier, N., Nickel, C., Zimmermann, P.-C., Ackermann, S., Woike, J. K., et al. (2013). Correlates of diagnostic accuracy in patients with nonspecific complaints. *Medical Decision Making : An International Journal of the Society for Medical Decision Making, 33,* 533–543. doi:10.1177/0272989X12470975.

Hodges, B. (2013). Assessment in the post-psychometric era: Learning to love the subjective and collective. *Medical Teacher, 35,* 564–568. doi:10.3109/0142159X.2013.789134.

Jarjoura, D., Early, L., & Androulakakis, V. (2004). A multivariate generalizability model for clinical skills assessments. *Educational and Psychological Measurement, 64,* 22–39. doi:10.1177/0013164403258466.

Jones, P., Smith, R. W., & Talley, D. (2006). Developing test forms for small-scale achievement testing systems. In S. M. Downing & T. Haladyna (Eds.), *Handbook of test development* (pp. 487–525). New York, NY: L. Erlbaum Associates.

Kane, M. (1996). The precision of measurements. *Applied Measurement in Education, 9,* 355–379. doi:10.1207/s15324818ame0904_4.

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

Karelaia, N., & Hogarth, R. M. (2008). Determinants of linear judgment: A meta-analysis of lens model studies. *Psychological Bulletin, 134,* 404–426. doi:10.1037/0033-2909.134.3.404.

Kaufmann, E., & Athanasou, J. A. (2009). A meta-analysis of judgment achievement as defined by the lens model equation. *Swiss Journal of Psychology, 68,* 99–112. doi:10.1024/1421-0185.68.2.99.

Keller, L. A., Clauser, B. E., & Swanson, D. B. (2010). Using multivariate generalizability theory to assess the effect of content stratification on the reliability of a performance assessment. *Advances in Health Science Education, 15,* 717–733. doi:10.1007/s10459-010-9233-8.

Kotovsky, K., Hayes, J., & Simon, H. (1985). Why are some problems hard? Evidence from Tower of Hanoi. *Cognitive Psychology, 17,* 248–294. doi:10.1016/0010-0285(85)90009-X.

Kreiter, C. (2008). A comment on the continuing impact of case specificity. *Medical Education, 42,* 548–549. doi:10.1111/j.1365-2923.2008.03085.x.

Kreiter, C. D., & Bergus, G. R. (2007). Case specificity: Empirical phenomenon or measurement artifact? *Teaching and Learning in Medicine, 19,* 378–381. doi:10.1080/10401330701542776.

Larson, J. S., & Billeter, D. M. (2016). Adaptation and fallibility in experts' judgments of novice performers. *Journal of Experimental Psychology. Learning, Memory, and Cognition.*. doi:10.1037/xlm0000304.

Leight, K. A., & Ellis, H. C. (1981). Emotional mood states, strategies, and state-dependency in memory. *Journal of Verbal Learning and Verbal Behavior, 20,* 251–266. doi:10.1016/S0022-5371(81)90406-0.

Marcoulides, G. A. (1996). Estimating variance components in generalizability theory: The covariance structure analysis approach. *Structural Equation Modeling: A Multidisciplinary Journal, 3,* 290–299. doi:10.1080/10705519609540045.

Mattick, K., Dennis, I., Bradley, P., & Bligh, J. (2008). Content specificity: Is it the full story? Statistical modelling of a clinical skills examination. *Medical Education, 42,* 589–599. doi:10.1111/j.1365-2923.2008.03020.x.

McClelland, D. C. (1973). Testing for competence rather than for intelligence. *American Psychologist, 28*(1), 1–14.

Mellenbergh, G. J. (1996). Measurement precision in test score and item response models. *Psychological Methods, 1,* 293–299. doi:10.1037/1082-989X.1.3.293.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18,* 5–11. doi:10.3102/0013189X018002005.

Norcini, J., Anderson, B., Bollela, V., Burch, V., Costa, M. J., Duvivier, R., et al. (2011). Criteria for good assessment: Consensus statement and recommendations from the Ottawa 2010 Conference. *Medical Teacher, 33,* 206–214. doi:10.3109/0142159X.2011.551559.

Norman, G. R. (2008). The glass is a little full-of something: Revisiting the issue of content specificity of problem solving. *Medical Education, 42,* 549–551. doi:10.1111/j.1365-2923.2008.03096.x.

Norman, G., Bordage, G., Page, G., & Keane, D. (2006). How specific is case specificity? *Medical Education, 40,* 618–623. doi:10.1111/j.1365-2929.2006.02511.x.

Norman, G. R., Tugwell, P., Feightner, J. W., Muzzin, L. J., & Jacoby, L. L. (1985). Knowledge and clinical problem-solving. *Medical Education, 19*(5), 344–356.

Popham, W. J. (2009). Assessment literacy for teachers: Faddish or fundamental? *Theory into Practice, 48,* 4–11. doi:10.1080/00405840802577536.

Popham, W. J., & Husek, T. R. (1969). Implications of criterion-referenced measurement. *Journal of Educational Measurement, 6*(1), 1–9.

R Core Team. (2013). R: A language and environment for statistical computing. Vienna, Austria. http://www.R-project.org/.

Ray, A., & Wu, M. (2003). *PISA programme for international student assessment (PISA): PISA 2000 technical report*. Paris: OECD Publishing.

Richter Lagha, R. A., Boscardin, C., May, W., & Fung, C.-C. (2012). A comparison of two standard-setting approaches in high-stakes clinical performance assessment using generalizability theory. *Academic Medicine, 87,* 1077–1082. doi:10.1097/ACM.0b013e31825cea4b.

Ricketts, C., Freeman, A., Pagliuca, G., Coombes, L., & Archer, J. (2010). Difficult decisions for progress testing: How much and how often? *Medical Teacher, 32,* 513–515. doi:10.3109/0142159X.2010.485651.

Roberts, J., & Norman, G. (1990). Reliability and learning from the objective structured clinical examination. *Medical Education, 24,* 219–223. doi:10.1111/j.1365-2923.1990.tb00004.x.

Rutkowski, L., von Davier, M., & Rutkowski, D. (2013). *Handbook of International large-scale assessment: Background, technical issues, and methods of data analysis*. Boca Raton: Chapman and Hall/CRC.

Schuwirth, L. (2009). Is assessment of clinical reasoning still the Holy Grail? *Medical Education, 43,* 298–300. doi:10.1111/j.1365-2923.2009.03290.x.

Schuwirth, L. W. T., & van der Vleuten, C. P. M. (2006). A plea for new psychometric models in educational assessment. *Medical Education, 40,* 296–300. doi:10.1111/j.1365-2929.2006.02405.x.

Schuwirth, L. W. T., & van der Vleuten, C. P. (2011). Programmatic assessment: From assessment of learning to assessment for learning. *Medical Teacher, 33,* 478–485. doi:10.3109/0142159X.2011.565828.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30,* 215–232. doi:10.2307/1435044.

Shavelson, R. J., Ruiz-Primo, M. A., & Wiley, E. W. (1999). Note on sources of sampling variability in science performance assessments. *Journal of Educational Measurement, 36*(1), 61–71.

Sijtsma, K. (2006). Psychometrics in psychological research: Role model or partner in science? *Psychometrika, 71,* 451–455. doi:10.1007/s11336-006-1497-9.

Skrondal, A., & Rabe-Hesketh, S. (2007). Latent variable modelling: A survey. *Scandinavian Journal of Statistics, 34,* 712–745. doi:10.1111/j.1467-9469.2007.00573.x.

Slovic, P., & Lichtenstein, S. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance, 6,* 649–744. doi:10.1016/0030-5073(71)90033-X.

Swanson, D. B., Norman, G. R., & Linn, R. L. (1995). Performance-based assessment: Lessons from the health professions. *Educational Researcher, 24,* 5–11. doi:10.3102/0013189X024005005.

van der Vleuten, C. P. M. (2014). When I say … context specificity. *Medical Education, 48,* 234–235. doi:10.1111/medu.12263.

van der Vleuten, C. P. M., Schuwirth, L. W. T., Driessen, E. W., Govaerts, M. J. B., & Heeneman, S. (2014). 12 Tips for programmatic assessment. *Medical Teacher*. doi:10.3109/0142159X.2014.973388.

von Davier, M., Sinharay, S., Oranje, A., & Beaton, A. (2006). The statistical procedures used in national assessment of educational progress: Recent developments and future directions. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Psychometrics* (Vol. 26, pp. 1039–1055). Amsterdam: Elsevier.

Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2006). Reliability coefficients and generalizability theory. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Psychometrics* (pp. 81–124, Handbook of Statistics): Elsevier Science.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

Wimmers, P. F., & Fung, C.-C. (2008). The impact of case specificity and generalisable skills on clinical performance: A correlated traits–correlated methods approach. *Medical Education, 42,* 580–588. doi:10.1111/j.1365-2923.2008.03089.x.

Wimmers, P. F., Splinter, T. A., Hancock, G. R., & Schmidt, H. G. (2007). Clinical competence: General ability or case-specific? *Advances in Health Science Education, 12,* 299–314. doi:10.1007/s10459-006-9002-x.

Wrigley, W., van der Vleuten, C. P. M., Freeman, A., & Muijtjens, A. (2012). A systemic framework for the progress test: Strengths, constraints and issues: AMEE Guide No. 71. *Medical Teacher, 34,* 683–697. doi:10.3109/0142159X.2012.704437.

Zumbo, B. D. (2006). Validity: Foundational issues and statistical methodology. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Psychometrics* (pp. 45–80). Amsterdam: Elsevier.