Introduction

In high-stakes assessments in medical education, such as licensure exams, the decision to let a particular participant pass or fail has far-reaching consequences (Swanson and Roberts, 2016). In order to become medical doctors, candidates have to show that they possess the required knowledge as well as the clinical skills - they need to prove that they are fit for medical practice. As certified physicians, their everyday work will consist of making choices that immediately affect their patients' health. For instance, this includes critical decisions on diagnoses, treatments, and medications. At the same time, many of these decisions are made by individual physicians and patients and society trusts that doctors make these decisions carefully (Cate et al. 2010). Assessment, as a part of the various licensing procedures implemented, is one facet that ensures this trust – it determines who will and who will not be able to move on in their medical career. Hence, assessment is a 'high-stakes' situation for students, the health care system, and the society. Consequently, providing evidence for the trustworthiness of assessments in medical education is a key responsibility that medical schools and licensing boards share.

Providing evidence for the reproducibility of test scores has been the "s*ine qua non"* of sound educational and psychological measurement (Parkes 2007, p. 2); the necessary condition for defensibleness and trustworthiness of decisions based on assessment data. In the words of Norcini et al., reproducibility means that the "results of the assessment would be the same if repeated under similar circumstances." (Norcini et al. 2011, p. 210). Reproducibility is closely related to the psychometric concepts of measurement reliability and measurement precision. Conversely, the notion that a consequential decision might be influenced by measurement error, i.e., an influence unrelated to the assessed skill or capacity, is in stark contrast to the basic concept of justness and fairness. An erroneous assessment cannot guarantee a reproducible, defensible outcome. However, what constitutes an assessment's outcome, its 'results', is dependent on the inferences made from the assessment. For instance, in one context, results may be used to rank students according to performance. In a different context, results may be used to make pass-fail decisions for single individuals. Consequently, the procedures used to estimate the reproducibility of an assessment's results and the actual use of these results must be aligned.

As already noted, pass-fail decisions carry some of the most critical consequences in medical education. Typically, those decisions represent evaluations of each individual's proficiency. Such judgments on whether or not a given particular student is sufficiently competent is the main purpose in typical high stakes testing contexts in

medical education (Eva and Hodges 2012). In order to make such a judgment, individual test scores are compared to a purposefully set cut-score, which usually represents a minimum acceptable level of ability (for methods related to setting defensible cut-scores, see Cizek, 2012, or McKinley and Norcini, 2014). The procedure is simple: if a student's test score is above the cut-score, she passes the exam; if not, she fails. In this manner, students are classified as being sufficiently competent to proceed with their studies or enter medical practice. From a psychometric perspective, pass-fail decisions can be understood as an individual-level classification. Hence, the important statistic to provide when justifying a pass-fail decision is the reproducibility of this individual-level classification.

Reliability coefficients seem to be widely used as an argument for the defensibleness of the inferences made from assessments in medical education (Norcini et al. 2011, 2018). Indeed, many psychometric analyses of assessment procedures or exams report estimates of reproducibility that are based on Classical Test Theory (CTT) or Generalizability Theory (G Theory) (Brannick et al. 2011; Hays et al. 2008). However, it is important to note that reliability coefficients (and procedures based on these coefficients, such as in Hays et al., 2008) do not inform on the reproducibility or precision of classification decisions on an individual level. Kane (Kane 1996, p. 366) states that "neither reliability coefficients nor generalizability coefficients provide an adequate analysis of precision in those contexts in which tests are used to make classification decisions". Indeed, the use of traditional reliability coefficients as evidence for the defensibleness of inferences of individual competence is troubling. Solely relying on these coefficients leads to unwarranted conclusions about the level of precision of pass-fail decisions for individuals (Huynh 1990; Subkoviak 1976; Webb et al. 2006).

Despite concerns about the mismatch between typically used estimates of reproducibility and decisions actually made, most texts on psychometrics in the medical education literature either neglect to mention the necessity to provide estimates of the precision of pass-fail decisions for single individuals or address this necessity only briefly in introductory texts (Champlain 2010; Downing 2003; Hays et al. 2008; Norcini 1999; Pell et al. 2010; Schuwirth and van der Vleuten 2011; Tavakol and Dennick 2012, 2013) and in texts on quality criteria for assessment in medical education (Norcini et al. 2011). With this module, we intend to fill this gap. Furthermore, we argue that the approach presented here is aptly aligned to the use of test scores to arrive at pass-fail decisions. Consequently, this article has two specific objectives. First, we briefly point out why reliability coefficients are inappropriate to

determine the reproducibility of a pass-fail decision for a single individual. Second, we delineate how to quantify measurement precision for pass-fail decisions based on Item Response Theory (IRT).

We chose IRT deliberately as a guiding framework for addressing the issue at hand because it provides the basis for criterion referenced interpretations since both items and persons are put on the same scale, and therefore, are directly comparable. Furthermore, IRT is the natural choice for analyzing categorical response data. Moreover, we argued elsewhere that, from a conceptual point of view, it is a good fit for assessment in medical education (Schauber et al. 2017). Finally, it seems that IRT has become the *de facto* standard within the field of educational assessment. Illustrating the usefulness of IRT for assessment in medical education might help to foster understanding of some of the basic concepts in this approach. However, a careful delineation of the advantages and drawbacks of IRT as opposed to other psychometric approaches in the current context is beyond the scope of this paper.

Importantly, procedures used to estimate the precision of a pass-fail decision are usually referred to as methods for estimating a given test's classification accuracy. This issue has been reported on extensively in the broader literature on educational measurement and psychometrics (Huynh 1990; Kane 1996; Lathrop and Cheng 2014; Lee 2010; Lewis and Sheehan 1990; Subkoviak 1976; Rudner 2005; Webb et al. 2006; Wyse and Hao 2012). These works, and especially the approach described by Rudner (2005), form the psychometric background for the following illustrations. As we aim for a conceptual illustration, more technical elaborations are beyond the scope of this paper.

Pass-Fail Decisions and Reliability

Traditional reliability coefficients (e.g., Cronbach's Alpha) can be understood as a summary statistic of the "precision for a population of subjects" (Mellenbergh 1996, p. 293). Put differently, a high reliability coefficient would indicate that the order of a group of individuals – from, for example, the best performing to lowest performing - is quite stable over occasions and can be reproduced well. A very low coefficient would suggest that such an ordering greatly changes from occasion to occasion. In practice, reliability coefficients are appropriate whenever inferences are made on the group level, that is, when we look to determine the relative standing of an individual within and across groups. Such between-person differences are, for instance, of interest in studies that

compare different instructional approaches, in which a clinical reasoning assessment may be used as an outcome to evaluate the effect of a specific instructional approach on students' reasoning skills. The question posed in this context is whether or not the intervention has an effect on the entire group of students; in this case a reliability index such as Cronbach's Alpha gives crucial information, as it indicates to what degree the clinical reasoning assessment captures the effect of the intervention. It is only when the test is able to sufficiently differentiate between persons in the intervention and control groups that this outcome measure can reflect a difference in performances between intervention and control conditions.

However, conclusions drawn in the context of a pass-fail decision are clearly different from those of an intervention study. A high degree of reproducibility in the context of pass-fail decisions means that, if administered under similar circumstances, a particular student would have been assigned the same classification – a pass would stay a pass, a fail would stay a fail. Thus, both the inference and the consequence are on the individual level. In this context, the expected reproducibility of a pass/fail decision is related to two magnitudes: first, the distance between a specific test score and the cut-score; and second, the amount of measurement error for that particular test score. Both magnitudes combined can be used to estimate the reproducibility of a pass-fail decision for a single individual.

Measurement error relates to the fact that, as with any other measured magnitude, test scores are expected to fluctuate randomly, i.e., they are not equivalent across similar administrations. Clearly, the farther away test scores are from the cut-score, the smaller the possibility that such fluctuations affect pass-fail decisions. For example, if the cut-score for an exam is set at 60%, i.e., responding correctly 60% of the items, we would hardly expect that a student who received a test score of 95% would have failed the exam under slightly different conditions. This student's test score might differ – but probably not to such an extent that she would fail the exam. On the other hand, for a student with a test score of 61%, the pass decision might easily have turned out into a fail (<60% correct) even under highly similar conditions. It might merely take one different question, or a lapse while marking the answers, to arrive at a test score of 59% – and thus, to fail. From an overarching perspective, in the context of high-stakes assessment the question is whether or not measurement error for a particular score is decisive for the given pass-fail decision. However, traditional between-person reliability indices cannot answer this question - a single number cannot adequately reflect the varying degrees of uncertainty involved in making individual pass-fail decisions.

Pass-Fail Decisions as a Hypothesis Test

From a statistical point of view, a pass-fail decision is highly similar to conducting a hypothesis test. To illustrate this, Figure 1 presents four selected test scores (labelled as a, b, c, d in Figure 1a, upper graph) based on simulated exam data. The four test scores are given as percentage-correct scores of 88%, 62%, 54%, and 6%, respectively. Furthermore, we chose to set the minimum pass-score at 60% correct answers. Similar to the considerations above, the 88% and the 6% scores would be rather clear pass and fail decisions. However, for instance, the 62% correct result might be more difficult to decide on. The hypothesis test can be conducted in two alternative but equivalent ways: either by using confidence intervals or p-values.

In Figure 1a, for each test score the 95% confidence interval was computed by adding (upper limit) and subtracting (lower limit) 1.96 times the respective standard error obtained from an analysis based on Item Response Theory. The 95% confidence interval means that, when samples are repeatedly drawn and confidence intervals are constructed in this way 95% of all confidence intervals will contain the true score (Morey et al. 2016). However, usually we want to draw inferences from only one sample (i.e., one test administration). Hence, if we want to make use of a particular confidence interval from our sample, we need to – as Neyman (1941) emphasizes – *decide* to act as if this confidence interval actually contained the true score. In the example given in Figure 1a, the confidence intervals for test scores (b) and (c) enclose the cut-score. Consequently, we consider that the corresponding pass-fail decision is too ambiguous to be defensible.

A complementary approach to confidence intervals is to regard the pass-fail classification as a statistical hypothesis on whether or not a person's true score is different from the cut-score. The corresponding p-value can then be interpreted as in other hypothesis tests, i.e., how likely the result is under the assumption of the null hypothesis that the true score equals the cut-score. Figure 1b (lower graph) gives p-values as a function of sum scores (x-axis) and a specific cut-score (60%) and highlights scores (a), (b), (c), and (d), which were also employed to illustrate the use of confidence intervals. Figure 1b also shows that the distance to the cut-score is related to lower p-values; the further the distance between the observed score and the cut-score, the more unlikely it becomes to observe the corresponding result given that the true score equals the cut-score. Therefore, Figure 1b can also be understood as a (ir)reproducibility function that indicates how reproducible pass-fail decisions are expected to occur (i.e., higher p-values indicate lower reproducibility). As is common in the Neyman-Pearson tradition of hypothesis testing (Lehmann 1993), the decision of where to set the threshold for rejecting the null hypothesis is critical. In this

case, we opted for a level of $\alpha = .05$, which would also indicate the expected percentage of false-positive decisions in the long run (i.e., incorrect rejection of the null hypothesis). Scores (b) and (c) exceed this pre-defined threshold and therefore fall in an area we designated 'ambiguous decision', while scores (a) and (d) would be deemed sufficiently reproducible to make a conclusive and defensible judgment.

Measurement precision of pass-fail decisions from the perspective of Item Response Theory

In this section we will delineate an approach that can give a more appropriate evaluation of the expected reproducibility of a pass-fail decision within an IRT framework. Item Response Theory is, indeed, a very broad framework used for analyzing various types of data and is, for instance, employed in many international large-scale assessments such as the PISA studies. The main aim of IRT is to estimate a person's level of proficiency or ability given her or his responses to a certain exam or test and quantify the magnitude of error associated with this estimation. As results, IRT analyses provide a score that reflects a student's ability and the precision associated with that score. This precision helps us to understand how sure we really can and should be that this student failed or passed (see examples above). In the following, we describe and illustrate how to quantify the precision associated with individual pass-fail decisions. We briefly introduce important key concepts of IRT and highlight the key steps necessary to arrive at the kind of function given in Figure 1 from the observed item responses[1]. Readers interested in an accessible introduction to IRT within medical education may refer to Downing (2003) or De Champlain (2010). DeMars (2010) and Embretson and Reise (2000) give a more general treatise of IRT. For a treatment of the statistical foundations of the specific IRT models used in this paper, please refer to texts by Baker and Kim (Baker and Kim 2010) or in Fischer and Molenaar (1995).

Step 1: From item responses to IRT item parameters

In a Rasch model, success (or lack thereof) in answering an item correctly is conceived of as the outcome of a direct comparison between a test taker's ability and the difficulty of an item. These are the two necessary factors to determine the probability of success on that specific item. Usually, the first step in a Rasch-based analysis is to estimate the difficulty of all items combined. When designing an assessment, the professionals involved may sense

[1] In the present article, we used the Rasch model for our illustrations. Thus, all considerations specifically apply to this model. However, the utilized concept of measurement precision is universal and applies to all IRT models.

that certain items should be easier for more certain students and others may be deemed more difficult to answer. Measurement models are the statistical tools to quantify such intuitive assumptions and to define mathematical relations between items and persons. While the relationship between proficiency and likelihood of success is implicit in CTT, it is explicitly formulated in IRT. In order to establish this relationship, item and person parameters (difficulty, ability) are placed on the same continuous scale (in a Rasch model called logit scale, as values are logarithmized odds ratios). Establishing a common scale based on students' responses to items is the first step in any IRT analysis. Importantly, in Rasch-like models (i.e., with equal item discriminations), there is still a perfect correspondence between ability estimates, usually referred to as theta, θ, and the percent-correct scores, where one particular sum score corresponds to one particular value on the IRT scale. This relation allows investigators to place the %-correct cut-score on the ability scale (Figure 2).

When this common scale – subsequently referred to as the ability scale – is established, Item Characteristic Curves (ICC) can be employed to depict the relationship between the probabilities for a correct response on a particular item for any level of θ. An item's difficulty is, by definition, set at the level of θ for which the success probability is 50%. Figure 3a gives an ICC for an item with a difficulty of zero on the ability scale. The function itself characterizes how, for this item, θ is related to the probability of success; the higher a person's ability, the more likely he will answer that item correctly.

Step 2: From item parameters to item information

Once an item's difficulty is determined, the expected probability of a correct response given a certain level of ability can be derived from the ICC. In order to describe this information conceptually, Figure 3 gives examples in which a constant area of probabilities of success (the y-axis) is projected onto the ability scale (x-axis). As the relationship between θ and probability of correct response is nonlinear, these projected areas vary along the ability scale, while the margins for success probabilities remain constant. Close to the item's difficulty (i.e., the ICC's inflexion point), a 4%increase in probability for a correct response corresponds to an absolute difference on the ability scale of 0.2 (0.2 - 0.0 = 0.2) (Figure 3b). However, in the upper tail of the function, the same increase in probability (4% from 0.95 to 0.99) is associated with an absolute difference of 1.6 (4.5-2.9=1.6) on the ability scale (Figure 3c). Thus, if, for example, a number of items with identical ICCs were given to a student, and if that student responded correctly to about 98% of these items, her θ estimate might vary between 2.9 and 4.5. On the other hand, students with a true

θ between 2.9 and 4.5 might have an almost identical probability of success. Therefore, those items would not be very informative in that area of ability. At the same time, if a student responded correctly to about 50% of the items, his θ might vary between 0 and 0.2 (Figure 3b), which is -obviously a much narrower range of probable θs. In this example, the level of proficiency of the student with a score of 50% can be narrowed to a much more restricted range of ability and thus can be measured more precisely than for the student who responded correctly to 98% of the items. Because this inference from the (expected) probabilities to the ability level near the turning point has tighter margins, it is more informative.

In more mathematical terms, there is more information in the middle of the ICC because the expected variance of the estimate (θ) is smaller there. Put differently, values of θ that correspond to a specific range of probabilities of success are more similar near the turning point than in the tails of the curve. This concept is more generally referred to as Fisher information and can be summarized for every single item in the so-called Item Information Function (IIF). The IFF can be derived from the ICC rather easily by calculating the product of the chance to answer the item correctly and the chance to answer the item incorrectly for any point on the ability scale (i.e., the variance of the Bernoulli distribution) (Figure 4). The maximum information for one item in a 1-parameter-logistic model is $I_{item} = 0.25$ because, at the turning point, $prob_{(correct)} = prob_{(incorrect)} = 0.5$, therefore $I_{item} = 0.5 * 0.5 = 0.25$.

Step 3: From item information to test information

After item difficulties are established and item information is derived from these difficulties, the IIFs of all the items in a test can be summed. The summed IIFs constitute the Test Information Function (TIF), which is the basis for calculating the conditional standard errors of measurement across the ability scale. If, for example, all four items in a four-item test have the same difficulty parameter of $Diff_{(item1-4)} = 0$, the test provides the information $I_{(test)} = I_{item1} + I_{item2} + I_{item3} + I_{item4} = 0.25 + 0.25 + 0.25 + 0.25 = 1$ at an ability level of $\theta = 0$ (Figure 5a). Here, the TIF has its highest value of 1 at θ=0. Figure 5b gives a second example of four items with difficulties of *Diff* = -2, -2, 1, and 2, respectively. This different distribution of item characteristics leads to a different TIF shape. While the same amount of information is available in total, it is more evenly spread across the ability scale. Hence, the maximum of the TIF is lower in this second example. Importantly, because the TIF ultimately depends on the characteristics of the included items, the conditional standard errors of measurement – and therefore measurement

precision – may vary for different tests. This is a very important feature of IRT, as items can be selected purposefully to reach sufficient measurement precision where it is deemed most important.

Step 4: From test information to *p*-values

The TIF forms the basis for calculating conditional standard errors of measurement across the ability scale. In IRT, the standard error of θ is defined as $SE_{(\theta)} = \sqrt{\frac{1}{I}}$. In our four-item example, the information for a θ value of 0 is $I_{(\theta=0)}=1$, thus the according standard error at $\theta= 0$ is $SE_{(\theta= 0)} \sqrt{\frac{1}{1}} = 1$. The corresponding 95% confidence interval for that ability level would then be $CI_{(\theta =0)} = 0 \pm 1.96 * 1$. The standard errors derived from the TIF can also be utilized to conduct a statistical test on whether or not a particular score is different from the cut-score. Specifically, the null hypothesis of whether a student's true score (estimated by $\hat{\theta}$) is equal to the cut-score (H0: $\theta_{(true)} = \theta_{(cut)}$) can be tested. The undirected alternative hypothesis is that the true score (estimated by $\hat{\theta}$) is different from the cut-score (H1: $\theta_{(true)} \neq \theta_{(cut)}$). With this formulation, it is possible to calculate the probability that the obtained score estimate (or a more extreme one) will be observed assuming the null hypothesis is true. This probability (i.e., the one-sided p-value) represents the probability of values equal to, or greater/lower than, the test statistic $(\hat{\theta} - \theta_{cut}) / SE(\hat{\theta})$ (Wyse and Hao 2012) under the standard normal distribution. Furthermore, since this distribution is symmetric, the two-sided p-value for the undirected hypothesis test is obtained by doubling the one-sided p-value. If the p-value is sufficiently low (i.e., below a consented value of, e.g., $\alpha = .05$) the null hypothesis is rejected and the alternative hypothesis is assumed to be true instead. This is because a p-value of < .05 means that there is only a very small chance of observing a difference as large as (or larger than) the one found given that the null hypothesis is true. If the p-value is below the pre-defined threshold of 5%, we interpret this as having reached statistical significance, at which point we decide to believe that a person's true score is not equal to the cut-score. As is true for any kind of null hypothesis test, this conclusion may be wrong simply due to chance, but we would still decide that this as a sufficiently defensible basis on which to determine competence because our probability of falsely rejecting the null hypothesis (the type I error rate which corresponds to α) is with 5% quite small.

An Applied Scenario

The previous delineations were based on conceptual considerations and illustrated by synthetic data. In the following subsections, we demonstrate the proposed approach using results from an actual exam.

Educational context and data

We used data from a randomly selected end-of-term exam administered at the end of the second year in the medical training program at the Faculty of Medicine, University of Oslo. The exam consisted of 110 items (multiple choice, multiple responses, and short essay) and was completed by a total of 70 students. Due to local regulations, to pass this exam, students had to respond correctly to 65% of the items. In the Norwegian context, end-of-term exams are part of the general licensing process, since no general national licensing exam exists. Students can re-sit a particular exam three times. If they fail the third re-sit, they are forced to drop out of medical training and will not be able to practice as a physician.

Ethical approval

Consent for the use of anonymized exam data was given by the Norwegian Social Science Data Services, under the reference number 43166, in August 2015.

Statistical analyses

All data processing and analysis were conducted in the R Language for Statistical Computing (R Core Team 2016). We used the TAM package (Kiefer et al. 2017) to estimate item and person parameters and their corresponding standard errors. Since short essay items and multiple response items included partial credit, a Partial Credit Model was used. The analysis was conducted according to the steps previously outlined in this paper. In order to determine the cut-score on the IRT scale, we estimated θ values for all possible numbers of correct answers. The θ score closest to 65% was set as the cut-score on the IRT scale, against which test-taker's ability levels were compared.

Results

Based on a CTT analysis, average item difficulty was 80% correct (standard deviation [SD] 19.9) and ranged between 15% and 84%. The average percent-correct score was 80% (SD 5.9) and ranged between 63.7% and 91%. Cronbach's Alpha for this exam was 0.82. The so-called 'EAP reliability' (Adams 2005), interpreted analogously to CTT reliability, was 0.88.

For a selected number of students, Table 1 gives the results from the analysis. For instance, one student had a score of 88.2% correct (Person F, Table 1) and thus scored clearly above the cut-score set at 65% correct. The corresponding p-value was lower than 5%, indicating that he would also – statistically – be regarded a "clear pass": It is highly unlikely that this student would have scored that high if he was actually 'incompetent'. For another student, the pass-decision is not as clear: Person D scored just above the cut-score, hence the decision to let him or her pass is highly ambiguous. This is indicated by a p-value of 84%. Table 1 also gives the theta values estimated in the IRT analysis. They denote a person's proficiency on a special metric, often called "logit metric" or "Rasch metric" or "theta scale". Fortunately, there is a one-to-one correspondence between the percent correct scores and the theta values so that they can be transformed back and forth. For each theta, a conditional standard error of measurement is provided as a result of estimating the Rasch model. These standard errors reflect the uncertainty associated with measuring the students' proficiencies and can thus further be used to evaluate how sure we can be about whether or not a student's true proficiency is above or below the cut-score. A more accessible form of presenting such results is plotting p-values against percent-correct values as shown in Figure 6. In this figure, we see the functional dependency ("curve") of p-values depending on percentage correct scores and the cut-score. The actual test scores from individual students are marked by a point on the function. Figure 6 illustrates as well that scores close to the cut-score had a less credible pass/fail decision (indicated by higher p-values) compared to scores farther from the cut-score. We also highlighted the area of ambiguous decisions, that is, the range of scores with a p-value higher than 5%. Critically, although the overall exam had a level of reliability exceeding 0.80, a number of individual pass-fail decisions (7 out of 70) for this exam would still fall within the area of ambiguous decision. Thus, ultimate pass/fail decisions for these students cannot sensibly made, because the uncertainty of whether their true proficiencies are above or below the cut-score is too high.

Discussion

In this article, we highlighted the key problem of using reliability coefficients to justify the defensibleness of pass-fail decisions (and other classificatory decisions). Indeed, traditional between-person reliability is a summary statistic that informs on the average reproducibility of test results for groups. Therefore, reliability coefficients are

not suitable for the evaluation of pass-fail decisions for single individuals. This has been widely acknowledged in the psychometric literature. Importantly, Kane (1996) highlights that the quantification of measurement precision is dependent on the consequences of the test scores and is a critical part of any argument for the validity of the test results. Employing an IRT framework, we illustrated how to arrive at a more appropriate evaluation of measurement precision in a context where pass-fail decisions are made. Employing a real-data example, we highlighted that a sufficient level of reliability does not mean that all decisions are expected to be highly reproducible. The most critical decisions, i.e., those near the cut-score, were expected to be the least replicable.

However, the approach we propose has drawbacks. One of the concerns for using IRT in typical medical school assessment scenarios is that it might not be possible to reach the assumptions underlying the models nor the necessary number of cases. As our main aim was to highlight the practical value of and the need to use conditional standard errors of measurement to evaluate the reproducibility of pass-fail decisions, we did not address issues related to the evaluation of model assumptions. However, it must be noted that recommendations on the minimum number of cases per parameter vary markedly in the literature. For instance, it has been suggested that the Rasch model may appropriate for as little as 50 respondents (Jones et al. 2006). Such recommendations also vary by application. For instance, assumptions of dimensionality may be tested in sample sizes as small as 250 (de Champlain and Gessaroli 1998). Although this number is small in comparison to large-scale assessment programs like the United States Medical Licensing Examinations, it is not attainable for a number of medical schools. An important drawback is that the approach presented here may be difficult to implement by medical schools. It still seems to be a rather rare scenario that staff is trained and skilled in educational measurement and statistics enough to contribute to the quality assurance of exams in medical education.

Furthermore, IRT is often regarded as a psychometric approach that makes strong mathematical assumptions, and deviations from such assumptions are to be expected in any kind of modelling scenario. The critical question is to what degree such deviations interfere with the interpretation of the derived consequences. In the context given here, violations of assumptions may lead to an inappropriately estimated Test Information Function, which would, in turn, lead to inaccurate conditional standard of error measurements. Unfortunately, to-date there is very little advice on the conditions in which, or the degree to which, the TIF is robust against violations of the underlying mathematical assumptions. Therefore, the robustness of our proposed approach should be addressed in future studies.

One of the explicit choices in this application is that the α-value (i.e., the threshold for statistical significance) for evaluating pass-fail decisions was set at 5%, which is common in a null hypothesis significance testing. There are two issues related to this approach. First, setting such a threshold should be a carefully considered and adjusted in a way meaningful to the responsible decision makers. Second, the (Frequentist) statistical framework employed here has its own limitations, which are present in most applications that employ a null hypothesis significance test or confidence intervals. Indeed, there has been substantial debate about the use and misuse of p-values and confidence intervals (Gelman 2013; Hoekstra et al. 2014; Morey et al. 2016). In this respect, Bayesian approaches are often considered to be more aligned with how researchers use and interpret the results of hypothesis tests. Importantly, this concern is not specific to the main issue raised in the current paper. Our argument is that estimates of precision need to be aligned with the level on which inferences are made, and accomplishing this is possible in both Bayesian and Frequentist frameworks.

Aside from these technical considerations, the proposed approach also raises policy-related issues. For instance, the functions presented Figures 1 and 6 ultimately indicate that, for pass-fail decisions, an area of ambiguity exists where neither a pass nor a fail decision seems to be justified. In fact, the most appropriate course of action from a psychometric view would be to obtain further information on students that fall into this area and delay the ultimate pass-fail decision until sufficient information is available. Right now, this seems to be a rather difficult endeavor, but there are other options. Students that fall into the area of ambiguous decision could be given the benefit of the doubt. That is, students below the cut-score but within the area of ambiguous decision might be let through because the assessment could not unambiguously ‘prove’ their inability. On the other hand, from a health care system perspective, students in the area of ambiguous decision that scored above the cut-score could be placed in the fail category, as their possible lack of competence could be regarded as a threat to patient safety. Again, such considerations are more policy-related and not specifically psychometric in nature. The procedure illustrated here offers a decision criterion that can be systematically applied, and a quantification of corresponding uncertainty that can be used to justify the defensibleness of pass-fail decisions for single individuals. We must carefully consider how to deal with the uncertainty and ambiguity in making high stakes decisions within a specific context and bear in mind the needs and demands of all stake-holders.

In conclusion, we hope to have raised awareness regarding an issue that is critical to the defensibleness of high-stakes decision made in assessments in medical education. The most critical point is that traditional between-person reliability coefficients are never appropriate as an argument for the defensibleness of a specific, individual, pass-fail decision. This issue has received very little attention in the literature on the use of psychometric procedures in assessment in medical education. Although there are several remaining issues, the approach we describe is one way to arrive at a more apt evaluation of measurement precision for high-stakes decisions for single individuals. At the very least, our considerations point out that relying on high reliability coefficients might be ill-advised, as they can lend to a contentedness that is inappropriate with regard to the defensibleness of critical decisions.

FIGURES

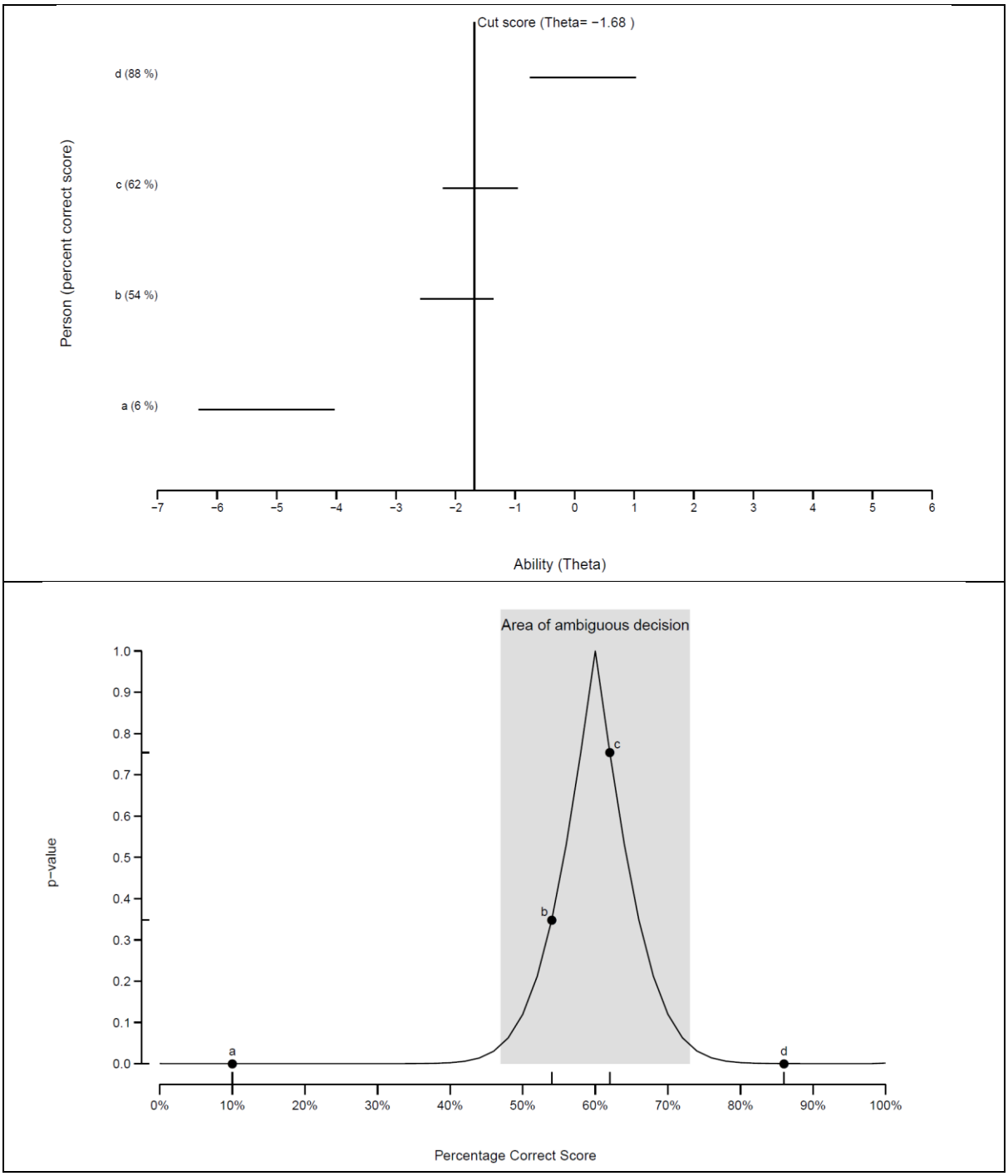

Figure 1a-b. The upper graph gives confidence intervals for four selected scores on the θ scale and their corresponding percent-correct score. The cut-score is marked as a vertical line. The lower graph gives a function that covers the (possible) observed scores on the x-axis and gives the according probability for a wrong pass-fail decision in form of a p-value (y-axis). The area of ambiguous decision is marked in grey. P-values of scores that fall in this area are above a pre-set significance level (α = .05).

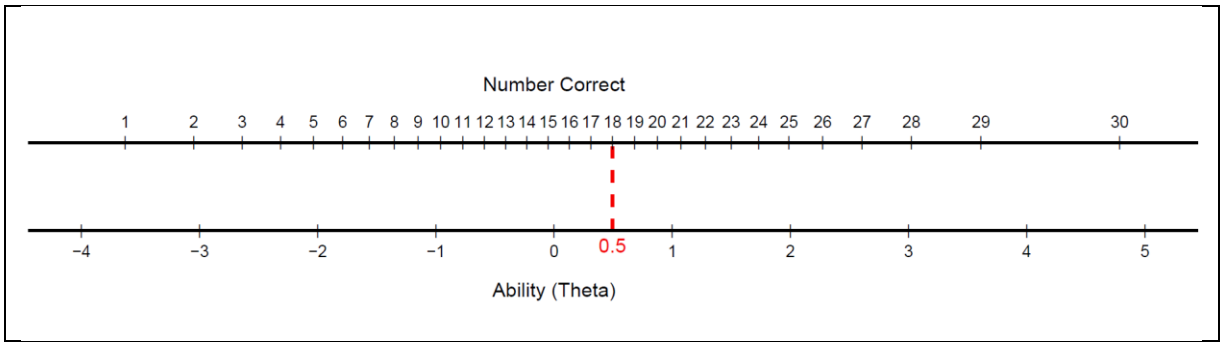

Figure 2. Illustration of the (nonlinear) relationship between ability scale (θ) and number-correct score scale for a simulated data set, adapted from DeMars (DeMars 2010). Since there is a direct correspondence between number-correct scores and the scores on the ability scale, a pre-defined cut-score (here: 18 correct answers) can be projected to the ability scale (θ = 0.5).

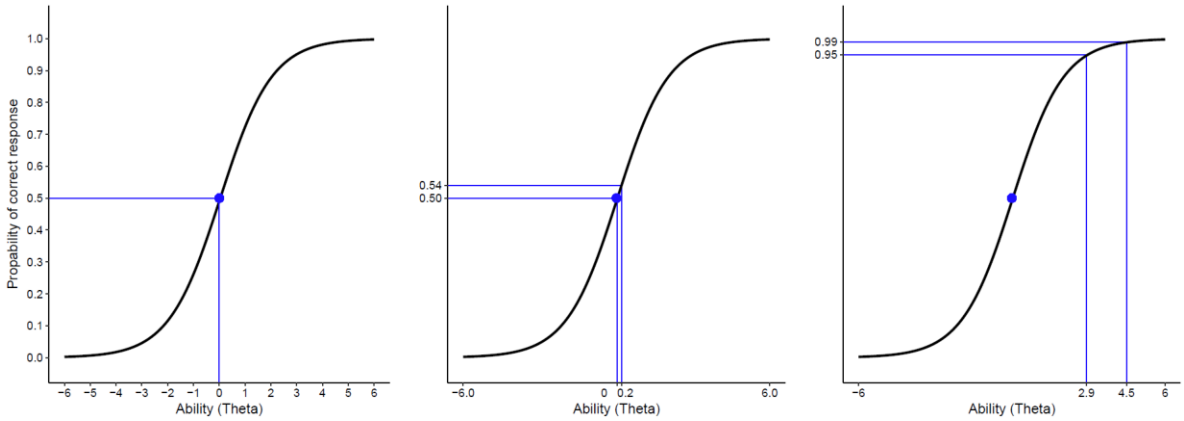

Figure 3a-c. The graph on the left depicts an estimated Item Characteristic Curve (ICC) with a location parameter of 0. Information can be regarded as a "projection" of an interval on the y-axis (probability of success) to a corresponding interval on the x-axis. This is illustrated in the middle and the right graph, where the closer the success probability gets to the inflexion point of the ICC, the higher the margins of the projection.

360

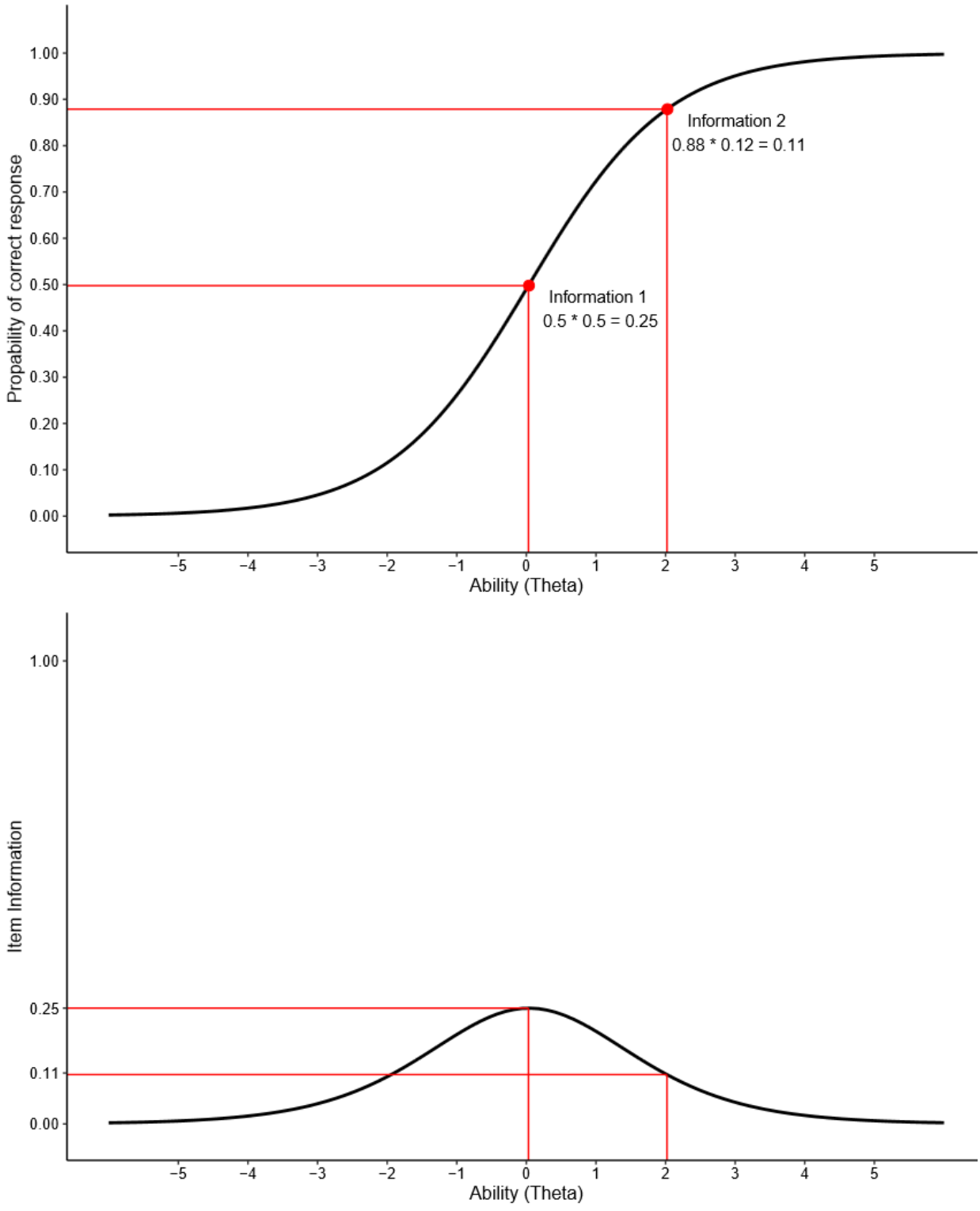

Figure 4. Item Characteristic Curve (upper graph) with two examples for the calculation of information at two ability levels, signified as Information 1 and Information 2 in the upper graph. The lower graph gives the Item Information Function for this item.

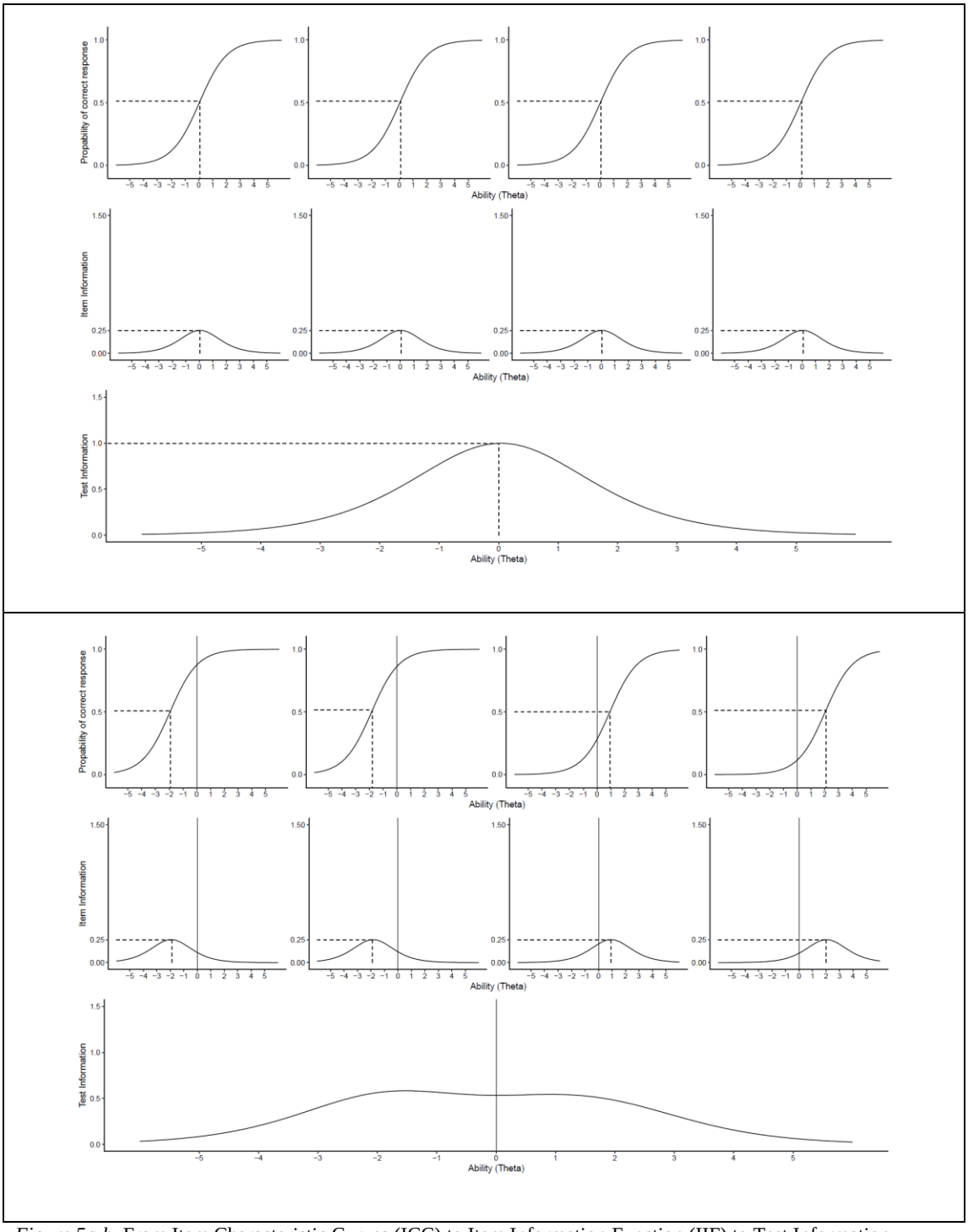

Figure 5a-b. From Item Characteristic Curves (ICC) to Item Information Function (IIF) to Test Information Function (TIF) for different item difficulties. The upper graph shows ICCs, IFFs, and the TIF derived for four items with a difficulty of zero. The lower graph illustrates how different item parameters lead to a differently shaped TIF. Here, item parameters are set to a difficulty of -2, -2, 1, and 2 for the four items included in the exam, respectively.

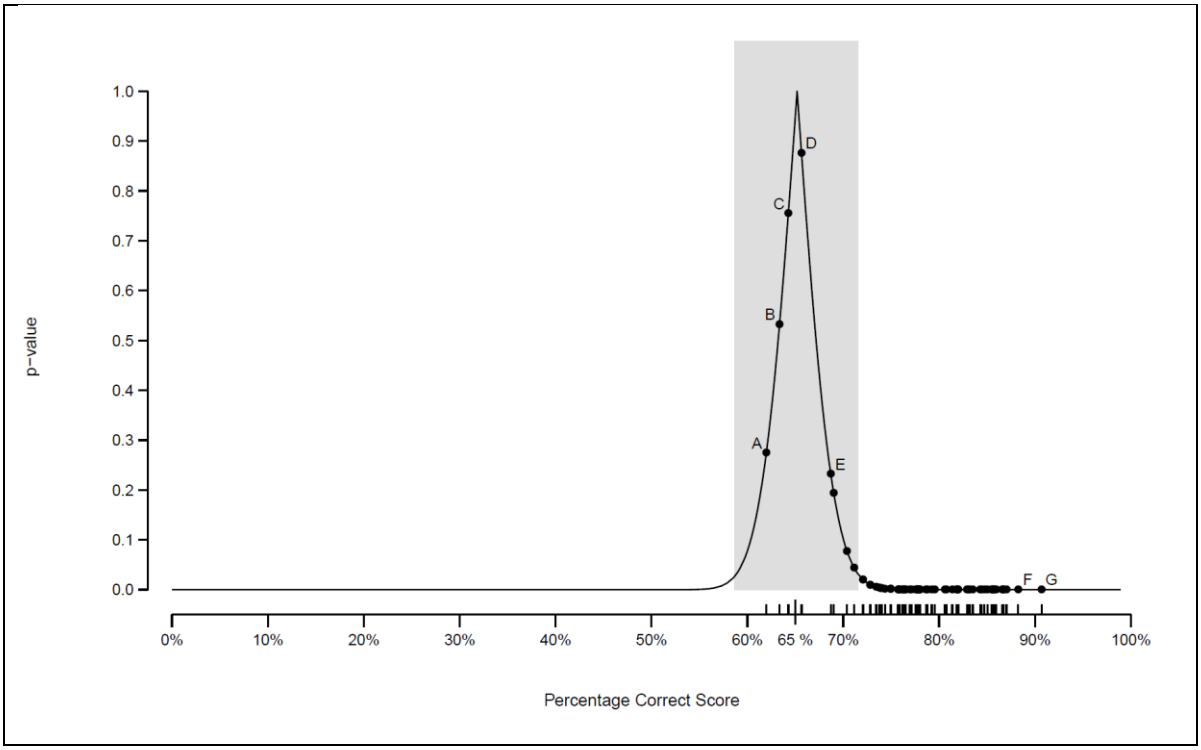

Figure 6. Estimated reproducibility function for a real-data example. Seven selected scores are marked from A to G and are also given in Table 1. While decisions for scores F and G are unambiguous (indicated by p-values below the pre-defined threshold), scores A to E fall into an area of ambiguous decision.

TABLES

Table 1: Selected results for seven students from an actual exam

Person	Percent-correct score	θ	cSEM	p-value
A	62.0	-0.29	0.05	.30
B	63.4	-0.27	0.05	.57
C	64.3	-0.25	0.05	.79
D	65.7	-0.23	0.05	.84
E	68.7	-0.17	0.05	.21
F	88.2	0.30	0.08	< .05
G	90.7	0.41	0.09	< .05

Note. Cut-score was set at θ = -0.24 which corresponds to 65% correct answers. cSEM is the conditional standard error of measurement derived from an IRT model.

References

Adams, R. J. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation, 31*, 162–172 (2005). doi:10.1016/j.stueduc.2005.05.008

Baker, F. B., & Kim, S.-H. (2010). *Item response theory: Parameter estimation techniques* (2nd ed., Statistics, Vol. 176). New York, NY [u.a.]: Dekker.

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 (2011). doi:10.1111/j.1365-2923.2011.04075.x

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

Champlain, A. de, & Gessaroli, M. E. (1998). Assessing the Dimensionality of Item Response Matrices with Small Sample Sizes and Short Test Lengths. *Applied Measurement in Education, 11*, 231–253 (1998). doi:10.1207/s15324818ame1103_2

Champlain, A. F. de. (2010). A primer on classical test theory and item response theory for assessments in medical education. *Medical education, 44*, 109–117 (2010). doi:10.1111/j.1365-2923.2009.03425.x

Cizek, G. J. (2012). *Setting performance standards: Foundations, methods, and innovations / Gregory J. Cizek, editor* (2nd ed.). New York: Routledge.

DeMars, C. (2010). *Item response theory* (Series in understanding statistics. Measurement). New York: Oxford University Press.

Downing, S. M. (2003). Item response theory: Applications of modern test theory in medical education. *Medical Education, 37*, 739–745 (2003). doi:10.1046/j.1365-2923.2003.01587.x

Embretson, S. E., & Reise, S. (2000). *Psychometric methods: Item response theory for psychologists* (Multivariate applications). Mahwah, N.J.: Lawrence Erlbaum Associates, Publishers.

Eva, K. W., & Hodges, B. D. (2012). Scylla or Charybdis? Can we navigate between objectification and judgement in assessment? *Medical education, 46*, 914–919 (2012). doi:10.1111/j.1365-2923.2012.04310.x

Fischer, G. H., & Molenaar, I. W. (1995). *Rasch Models*. New York, NY: Springer New York.

Gelman, A. (2013). P values and statistical practice. *Epidemiology (Cambridge, Mass.), 24*, 69–72 (2013). doi:10.1097/EDE.0b013e31827886f7

Hays, R., Gupta, T. S., & Veitch, J. (2008). The practical value of the standard error of measurement in borderline pass/fail decisions. *Medical education, 42*, 810–815 (2008). doi:10.1111/j.1365-2923.2008.03103.x

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic bulletin & review, 21*, 1157–1164 (2014). doi:10.3758/s13423-013-0572-3

Huynh, H. (1990). Computation and Statistical Inference for Decision Consistency Indexes Based on the Rasch Model. *Journal of Educational and Behavioral Statistics, 15*, 353–368 (1990). doi:10.3102/10769986015004353

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 (1996). doi:10.1207/s15324818ame0904_4

Kiefer, T., Robitzsch, A., & Wu, M. (2017). TAM: Test Analysis Modules (1st ed.). https://CRAN.R-project.org/package=TAM.

Lathrop, Q. N., & Cheng, Y. (2014). A Nonparametric Approach to Estimate Classification Accuracy and Consistency. *Journal of Educational Measurement, 51*, 318–334 (2014). doi:10.1111/jedm.12048

Lee, W.-C. (2010). Classification Consistency and Accuracy for Complex Assessments Using Item Response Theory. *Journal of Educational Measurement, 47*, 1–17 (2010). doi:10.1111/j.1745-3984.2009.00096.x

Lehmann, E. L. (1993). The Fisher, Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two? *Journal of the American Statistical Association, 88*, 1242 (1993). doi:10.2307/2291263

Lewis, C., & Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. *ETS Research Report Series, 1990*, i-48 (1990). doi:10.1002/j.2333-8504.1990.tb01364.x

McKinley, D. W., & Norcini, J. J. (2014). How to set standards on performance-based examinations: AMEE Guide No. 85. *Medical teacher, 36*, 97–110 (2014). doi:10.3109/0142159X.2013.853119

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

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic bulletin & review, 23*, 103–123 (2016). doi:10.3758/s13423-015-0947-8

Neyman, J. (1941). Fiducial Argument and the Theory of Confidence Intervals. *Biometrika, 32*, 128 (1941). doi:10.2307/2332207

Norcini, J. (1999). Standards and reliability in evaluation: when rules of thumb don't apply. *Academic medicine : journal of the Association of American Medical Colleges, 74*(10), 1088–1090.

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 (2011). doi:10.3109/0142159X.2011.551559

Norcini, J., Anderson, M. B., Bollela, V., Burch, V., Costa, M. J., Duvivier, R., ... & Swanson, D. (2018). 2018 Consensus framework for good assessment. Medical teacher, 40(11), 1102-1109.

Parkes, J. (2007). Reliability as Argument. *Educational Measurement: Issues and Practice, 26*, 2–10 (2007). doi:10.1111/j.1745-3992.2007.00103.x

Pell, G., Fuller, R., Homer, M., & Roberts, T. (2010). How to measure the quality of the OSCE: A review of metrics - AMEE guide no. 49. *Medical teacher, 32*, 802–811 (2010). doi:10.3109/0142159X.2010.507716

R Core Team. (2016). R: A Language and Environment for Statistical Computing. Vienna, Austria. https://www.R-project.org/.

Rudner, L. M. (2005). Expected classification accuracy. *Practical Assessment Research & Evaluation, 10*(13), 1–4.

Schauber, S. K., Hecht, M., & Nouns, Z. M. (2017). Why assessment in medical education needs a solid foundation in modern test theory. *Advances in health sciences education : theory and practice*. doi:10.1007/s10459-017-9771-4

Schuwirth, L. W. T., & van der Vleuten, C. P. M. (2011). General overview of the theories used in assessment: AMEE Guide No. 57. *Medical teacher, 33*, 783–797 (2011). doi:10.3109/0142159X.2011.611022

Subkoviak, M. J. (1976). Estimating Reliability from a Single Administration of a Criterion-Referenced Test. *Journal of Educational Measurement, 13*(4), 265–276.

Swanson, D. B., & Roberts, T. E. (2016). Trends in national licensing examinations in medicine. *Medical education, 50*, 101–114 (2016). doi:10.1111/medu.12810

Tavakol, M., & Dennick, R. (2012). Post-examination interpretation of objective test data: monitoring and improving the quality of high-stakes examinations: AMEE Guide No. 66. *Medical teacher, 34*, e161-75 (2012). doi:10.3109/0142159X.2012.651178

Tavakol, M., & Dennick, R. (2013). Psychometric evaluation of a knowledge based examination using Rasch analysis: an illustrative guide: AMEE guide no. 72. *Medical teacher, 35*, e838-48 (2013). doi:10.3109/0142159X.2012.737488

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.

Wyse, A. E., & Hao, S. (2012). An Evaluation of Item Response Theory Classification Accuracy and Consistency Indices. *Applied Psychological Measurement, 36*, 602–624 (2012). doi:10.1177/0146621612451522