1

2 **Title**

3 Challenging the norm: Length of exams determined by classification accuracy or reliability

4

5

6 **Authors**

7 Stefan K. Schauber[1, 2] & Matt Homer[3]

8

9 **Affiliation**

10 [1]Section for Health Sciences Education (HELP), Faculty of Medicine, University of Oslo

11 [2]Centre for Educational Measurement (CEMO), CREATE, Faculty of Educational Sciences,
12 University of Oslo

13 [3]Leeds Institute of Medical Education, School of Medicine, University of Leeds, Leeds, UK

14

15 **Manuscript: 6507 words**

16 **Abstract: 295 words**

17

ABSTRACT

Purpose

This paper challenges the notion that reliability indices are appropriate for informing test length in exams in medical education, where the focus is on ensuring defensible pass-fail decisions. Instead, we argue that using classification accuracy instead better suited to the purpose of exams in these cases. We show empirically, using resampled test data from a range of undergraduate knowledge exams, that this is indeed the case. More specifically, we address the hypothesis that use of classification accuracy results in recommending shorter test lengths as compared to when using reliability.

Method

We analyzed data from previous exams from both pre-clinical and clinical phases of undergraduate medical education. We used a re-sampling procedure in which both the cut-score and test length of repeatedly generated synthetic exams were varied systematically. N=52,500 datasets were generated from the original exams. For each of these both reliability and classification accuracy indices were estimated.

Result

Results indicate that only classification accuracy, not reliability, varies in relation to the cut-score for pass-fail decisions. Furthermore, reliability and classification accuracy are differently related to test length. Optimal test length to using reliability was around 100 items, independent of pass-rates. For classification accuracy, recommendations are less generic. For exams with a small percentage of fail decisions (i.e., 5% or less), an item size of 50 did, on average, achieve an accuracy of 95% correct classifications.

Conclusions

We suggest a move towards the employment of classification accuracy using existing tools, whilst still using reliability as a complement. Benefits of re-thinking current test design practice include minimizing the burden of assessment on candidates and test developers. Item writers could focus of developing fewer, but higher quality, items. Finally, we stress the need to consider effects of the balance false positive and false negative decisions in pass/fail classifications.

INTRODUCTION

How to design, build, and deliver high-quality exams has been a long-term focus within the field of assessment in health professions education (HPE) [1]. Most faculty members would probably agree that designing high quality content for assessments—even for a 'simple' knowledge test, let alone across a programme of assessment—demands a significant investment of resources: multiple choice exams need appropriate clinical vignettes with reasonable distractors, and OSCEs need high quality stations for defensible decision-making.

An ongoing challenge is to decide on *how many* of these items, cases, or observations are enough for an exam—it is clearly more feasible to develop twenty carefully crafted case scenarios than one hundred [2,3]. A need for a greater quantity of items, stations, and assessments in general might also impact the quality of the content provided, at least in educational contexts where resources are typically limited. Hence, the question of how many observations (i.e., items, stations, cases, etc.) are required for good assessment remains. In investigating this question, we propose that educators and assessment scholars should re-consider the norms of psychometric practice for exams in health professions education.

From a global perspective, for many medical schools, decisions about 'how much is enough' in an assessment might often be informal, based on rules-of-thumb, or common practice at other schools, or even what the regulator 'expects'. However, for some high-stakes assessments, such as some national licensing exams, psychometric calculations can help to determine the number of observations needed as a prerequisite for defensible decisions [4–7]. More generally, in our experience and given examples in the literature, we find that a typical end-of-term exam in medical schools have of the order of 100 items, while licensing exams might

double or even triple that number. Again, providing that much exam content, especially to a high level of quality, and often multiple times per year, can be a challenging and time-consuming task for test developers. The need to provide extensive (i.e. sufficiently long) exams seems to be common in thinking on assessment: "The message here is that more data equates to a better picture" [8]. This notion suggests is that we need 'a lot' and that 'more is better', mirroring the idea of an increasing pixel resolution of an image [9].

In this study we aim to gently challenge this rationale, based on an underlying and nuanced psychometric perspective. We ground our study in the idea that the purpose of a given exam should be guiding the answer to 'how much is enough', and that the rationale applied for determining this 'how much' must fit that purpose. This is especially relevant in competency-based medical education [10] where there is an explicit focus on achieving particular outcomes. In brief, we will argue that the commonly employed statistic to determine test length—reliability—is an inferior, or even inappropriate, psychometric indicator for this purpose in many assessment contexts. We propose that focusing on an appropriate measure—classification accuracy [11,12]—can lead to the re-thinking and re-designing of common assessment practices. In simple terms, classification accuracy can be thought of as a model-based estimate of the proportion of candidates in an exam who are correctly classified as true passes or true fails.

To focus our arguments, we concentrate our empirical work on arguably the most straightforward and most common type of assessment in HPE, the standardized, written examination. However, we claim that our argument and the according psychometric rationale are applicable to other types of assessments, too, including performance assessments such as the OSCE, but also more complex assessment scenarios such as in programmes of assessment.

The traditional perspective on reliability and test length

Current frameworks regarding the quality of assessment include several important aspects, such as its educational impact as well as the extent to which it is a catalyst for improving learning and instruction [13,14]. In high-stakes testing, defensibility of the decisions made from exam scores is an integral part of assessment quality [13,15]. Here, two fundamental criteria for good assessment are that it is both reliable and valid. Critically, test length as a basic feature of any assessment is related to both aspects. In terms of validity theory[16], shorter tests might increase the risk of construct underrepresentation[17]. This means that an exam or assessment might fail to cover important aspects of the competency or skill that it tries to assess. With longer tests, the assessed content area can be covered more adequately[16,17]. In addition, longer exams tend to be more reliable[17,18], and determining the appropriate length of an exam is crucial in assuring good assessment. There is, indeed, a body of literature on determining the amount of content for a given assessment, as, for instance for credential exams where careful task analyses and weighting procedures are employed before specifying an exam blueprint[19,20]. However, there is little specific recommendation on how *many* items are needed to cover a domain appropriately, especially in complex and multi-faceted disciplines such as medicine. A common recommendation then is to sample broadly [21] or to revert to reliability-based calculations of appropriate test length[17,18]. Still, there remains little specific guidance on what constitutes enough observations, or enough items.

The number of items in an exam need not be completely arbitrary, and decisions on appropriate test length can be informed by psychometric analyses of real test data. The simplest application of such an analysis can be done using the Spearman-Brown formula [22], which, from real test data, estimates the number of items needed in a test to achieve a sufficient level of reliability. As argued earlier, from our perspective, the issue here is that reliability—a measure of

measurement precision—should be regarded as an indicator of lesser interest in many assessment contexts, particularly in comparison to classification accuracy. We develop these arguments in the next section.

The problem: reliability is not the same as accurate classification

Reliability can be conceptualized as a signal-to-noise ratio, that is, it provides an estimate of how much error there is in the scores obtained from an assessment [22,23]. Higher reliability indicates more measurement precision and less error. This property is important if each individual score matters equally, or if we are interested in the stability of the rank-order of candidates. For instance, this is usually the case in progress testing where the goal is to follow individual students' learning progression over time [24,25]. This was also the case in the previous USMLE-Step 1 where candidates received score reports, which, in turn, could be used for selection to further training [26]. In these cases, the purpose and the use of exam scores are well-aligned with the notion of reliability as an appropriate measure of psychometric quality.

But not all scores are equally important—at least not in general. A key scenario where different scores matter differently is when the main purpose of an assessment is to decide on each candidate's readiness for the next step of training, which is the case in many licensing examinations globally[27–29]. In these contexts, classificatory decisions are made using the score of an individual student: the final decision awarded is either a pass or a fail—such as has been recently implemented in the USMLE Step 1 [26] or the GMC's national licensing assessment in the UK. The key inference made from the test score is if the candidate is sufficiently competent, if they reach a "threshold for safe practice"[30]—or not.

For illustrative purposes, let us consider a hypothetical scenario where a pass/fail decision for every candidate is made. In a regular medical school, this might be based, say, on a

100-item multiple-choice exam. Furthermore, assume the passing score was defensibly set to 60% correct out of the 100 questions. We find a candidate who responded correctly to 90 out of the 100 questions will clearly pass the exam. Yet, for making the pass/fail decision, we are not particularly interested if their particular result—an observed score of 90—is measured with great precision. Neither are we interested in whether one candidate outperforms another. Rather, we want to know whether this candidate could have passed by chance, that is, if they were so lucky that, while in truth being not ready to progress, they passed with sufficient help of measurement error in their favor. This is highly unlikely in this case—both intuitively and psychometrically—given how far above the cut-score (60) their observed score is (90). While often very useful, our key point is that reliability indices tell us very little about the precision of the pass/fail decisions.

When we focus on classification decisions made based on exam outcomes, the notion of measurement error has a different interpretation to that associated with score reliability. If the purpose of an exam is to make pass/fail decisions, the spotlight is on a very specific issue. We must ask how many of the candidates in a particular exam are misclassified—deemed competent when they should fail and failing while they should have progressed. This is a question that a reliability coefficient does not answer—except in the case when there is no measurement error and, consequently, reliability "pushes 1" [31]. Thus, when classification decisions are to be made, such as in competent/not ready, what is most important is the accuracy of that classification[32,33], and not the amount of noise in each single score [34]. From a validity perspective, we argue that the inferences and intended use of the outcomes of an exam[16] in this case match much better with classification accuracy[35] than with reliability.

The general concept of classification is also in line with considerations of false positive and false negative decisions, that is, with the errors associated with passing truly incompetent

candidates and failing truly competent ones. In high stakes settings such as in licensing examinations balancing between false positives and false negatives is an important policy consideration. If "false positive" candidates—those deemed competent but in truth are not—progress into medical practice they pose a threat to patient safety. On the other hand, if a truly competent students fails an exam, they are likely to have a chance to resit the examination. The societal cost of the former typically outweighs the cost of the consequences of the latter. Although not the focus of this article, a focus on classification-accuracy stresses the need to be explicit about such policies.

Approaches to estimating classification accuracy have been developed under different psychometric frameworks and are well-known in the broader literature on educational measurement for decades [11,12,36]. While estimates of both classification accuracy and reliability provide important psychometric information in many contexts, to the best of our knowledge, there is no published study that aims at using classification accuracy as a guiding index for informing the design of high-stakes assessments in HPE where there is a specific focus on decisions such as competent or not-ready.

Understanding the difference between reliability and classification accuracy

In developing our understanding of the difference between reliability and classification accuracy, it is important to note that there is no role for the cut-score in the calculation of reliability coefficients. By contrast, estimates of classification accuracy include both candidates' scores and their distance to the cut score explicitly[11,12,33]. To further illustrate, imagine an educational context where assessment, learning, and instruction are very-well aligned. Here, all candidates are proficient learners, who went through effective instruction conducted by competent educators. After such a course, they take an end-of-term-exam. In such an idealized

context, students will likely score homogenously and distinctly above the cut-score: All students are competent regarding the material they are supposed to master. Psychometrically, however, this results in a low variance in scores, which in turn implies that reliability estimates might drop below recommended standards, potentially even approaching zero. Reliability measures in this context convey no information about how well the examination is sorting the candidates correctly or otherwise into passing or failing.

In the same idealized context, the focus on measures of classification accuracy is much more meaningful. Again, classification accuracy considers both candidates' scores and their distance to the cut-score. As all students are clearly competent, their scores exceed the pass-mark by some margin. Critically, the corresponding classification accuracy estimates will, at least theoretically, not be affected by the very distinct pattern of scoring and might even approach perfect accuracy in this example. Thus, classification accuracy is better aligned to the key purpose of this type of assessment.

To summarize our arguments, a simple but consequential aspect when designing an assessment is deciding on the number of observations needed to ensure defensibility [6,13,21]. From a broad validity perspective, there is a lack of specific guidance on "how much is enough" to ensure appropriate content coverage. From a test-design viewpoint, this issue remains an ongoing challenge and is one we cannot cover extensively in this paper. However, psychometric calculations can be readily applied to ensure appropriate measurement precision at a particular test-length, but these are usually based on reliability considerations. We argue that this approach is ill-aligned to the purpose of assessments in which classificatory decisions, such as competent/not ready, matter most. Hence, we propose to compare appropriate test-lengths for

assessments based on classification accuracy for pass/fail decisions with those based on more traditional reliability estimates.

Research Question

We hypothesize that when the purpose of an assessment is to classify candidates into competent or not, calculations based on classification accuracy will likely result in less extensive test lengths as compared to calculations based on reliability. To address this question, we use a range of real undergraduate medical assessment data to investigate the extent to which measures of classification accuracy and reliability arrive at different conclusions for optimal test length in different scenarios. These analyses have both theoretical and practical consequences. Theoretically, we contribute by highlighting the important differences in the estimates of measurement precision we use to provide evidence for the defensibleness of our assessment decisions. Practically, this work also has potentially significant implications for how exams could and should be designed in HPE and beyond.

We continue this paper outlining our methods and the results. We then situate our findings in the current literature and discuss what they might mean for test development in HPE. This will include the potentially challenging issue of ensuring sufficient domain sampling in complex domains such as applied clinical knowledge.

METHODS

In this study, we compare test lengths that either optimize reliability or measures of classification accuracy (CA). We use multiple administrations of three different knowledge tests from an undergraduate medical program. We detail in turn the data sample, ethical issues, our

analytic approach, as well as describing the measures of reliability and classification accuracy chosen.

Sample

Data analyzed here are retrieved from previous exams from the medical programme at the Faculty of Medicine at the University of Oslo. These are high-stakes end-of-term applied knowledge examinations. Students must pass these to continue with their studies. In our context, there is no national licensing exam comparable to the USMLE-Step-1 or Step 2/3 for the pre-clinical and clinical parts, respectively. Hence, the exams analyzed here are part of the general licensing procedure and their aim is to assure candidates' minimal competence via the award of an undergraduate medical degree.

The exam developments uses a cyclic system of quality assurance, which includes pre-exam proofing and post-exam item-level analysis [37]. For this study, we used students' response data from repeated administrations of one from the pre-clinical phase, and two from the clinical phase (so three exams and five administrations each = 15 exam datasets). These datasets include mostly single-best-answer multiple choice items, but also multiple response as well as short-essay questions. All exams covered the content taught in the relevant module. The pre-clinical exam covered anatomy, physiology, micro-biology and immunology. The dominant subjects included in the clinical exams were pediatrics, gynecology and obstetrics in one exam, and psychiatry and social medicine in the other one.

We only had access to anonymized data, which included students' scores on all items within an exam. Thus, for instance, an exam with 50 participating candidates and 100 multiple choice items could be presented in a spreadsheet of 50x100 item-level scores plus one anonymized candidate-ID variable. In the current context, there is a pre-defined fixed passing

standard and an elaborate pre- and post-exam quality assurance regime is put in place to maintain standards across administrations[37].

Ethics and informed consent

We stored and handled the data according University of Oslo's data policy for 'yellow' data [38], since these are exam data limited in volume and without any sensitive or person-identifiable information. We used the data without informed consent for the purpose of research in the public interest and within quality assurance in higher education. Each student's right to be informed about the data usage is accounted for by a public blog post on the project [39]. Permission to process the data was granted by the Norwegian Agency for Shared Services in Education and Research, reference number 497365.

Data sharing

Limited access to the student-level responses to exam items can be granted upon reasonable request to the first author (SKS). Results from the analyses are aggregate data which are available via a public repository[40].

Analytical approach

Scoring

The possible score per item per student varied between 0 and 100 percent correct. All items, regardless of response format, were weighted equally. Both student exam scores and cut-scores were calculated and handled as percentage-correct scores.

Resampling study

To address our research objective, we designed a resampling study [41], with resampling of items as the main approach. Put briefly, the procedure included repeated, independent draws of

item samples with replacement from the existing exam data described above. The number of candidates remained constant within each of the 15 exam datasets. The resampling study allows us to investigate how psychometric indices—reliability and classification accuracy —vary with other features of the assessment (i.e. number of items and cut scores). Consequently, the results will help to evaluate the usage of either classification accuracy and/or reliability as guides for designing exams for pass/fail decision-making.

We did not use the originally set cut-scores here since we did not have direct access to this information. Our approach is illustrative rather than attempting to replicate actual exam decision-making. Furthermore, and most importantly, we designed the study to show how psychometric indicators reflect the precision of classificatory decisions made in accordance with variation of assessment features. Hence, the actual estimates for the 'real' exams were of little interest in our research.

Overall, we varied two assessment features systematically. Firstly, we specified different test lengths, in accordance with previous research [42]. Secondly, we also varied the hypothetical cut-scores systematically. More specifically, we chose seven conditions for test length (20, 30, 50, 75, 100, 125, 150) and five distinct cut-scores (40, 50, 60, 70, 80), meaning that there were 35 possible combinations of test length and cut-score. For each for these 35 combinations we drew 100 random samples of items from each exam administration. Per exam administration, this means we drew 3500 samples and calculated reliability and classification accuracy for each sample. Since this was done for five previous administrations for the said three exams, we calculated a total of 3,500*3*5 = 52,500 estimates for reliability and classification accuracy.

All data handling, preparation and calculation of all relevant statistics were handled in the R Language for Statistical Programming [43]. Table 1 summaries our overall analytical approach in pseudo-code.

Measures of reliability and classification accuracy

As noted earlier, estimates of both reliability and classification accuracy are available within different psychometric frameworks. For instance, Cronbach's Alpha is a commonly used reliability estimate within a Classical Test Theory (CTT) framework, whilst Item Response Theory (IRT) offers estimates based on modelling of an underlying latent variable [44]. Similarly, classification accuracy can also be calculated within either CTT or IRT frameworks [11,12]. A more recent estimate for classification accuracy has been proposed by Lathrop & Cheng [33]. The interested reader may refer to these references for more technical details of the procedure. This is a non-parametric approach to classification accuracy which does not rest on strong statistical assumptions for the underlying exam data and thus is more flexible in its application.

In this study, for both classification accuracy and reliability, we calculated estimates based on CTT as well as IRT. We mainly focus on the following two indicators since they are well-established in the psychometric literature.

1.) Cronbach's Alpha as an estimate of score reliability
2.) Non-parametric classification accuracy using the methods proposed by Lathrop & Cheng[33]

However, to triangulate the results from these analyses, we additionally estimated IRT-based estimates of reliability and classification accuracy using procedures in the TAM[45] and cacIRT packages[46] respectively. Finally, we estimated classification accuracy based on the

Livingston and Lewis method using the betafunctions package [47]—which also provides an easily accessible front end freely available at https://hthaa.shinyapps.io/shinybeta/. In total, this gives two measures of reliability and three of classification accuracy per resampled exam.

While we, in this paper, focus on Cronbach's Alpha as an estimate for score reliability, we note that there are other frameworks for the calculation of estimates of measurement precision. Generalizability Theory, commonly used in medical education assessment research, has made important theoretical and analytical contributions to understanding and designing assessments. Scholars in this tradition have worked on topics closely related to making classificatory decisions. Here, particularly interesting is Kane's works on "tolerance for error"[48–50] while Brennan discusses estimates of measurement error that are meaningful for the context discussed here[51]. For the sake of brevity, we do not consider Generalizability Theory in further detail here.

Statistical analyses of resampling results

To investigate how the estimates of both reliability and classification accuracy relate to the two key features of the assessment (i.e., test length and the cut-score), we analyzed the results across the repeated samples and conditions using linear mixed effects models as implemented in the lme4 R-package [52]. For all models, the estimate of interest (either Cronbach's Alpha or classification accuracy) was the dependent variable. In terms of predictors, the models included fixed effects for test-length and the cut-score. We also included a random effect for the three exams to account for the fact that the repeated administrations of each of these exams are not entirely independent. The fixed effects were included in a stepwise procedure. Since we estimated a mixed effects model, we also report the intra class correlation (ICC) for the exam. In

our context, this ICC represents proportion of variance in the psychometric estimate that is exam-specific, that is, due to the clustering of repeated administrations within an exam.

Since the sample size for this part of the study was large (52,500), negligible effect sizes are likely to be statistically significant on the 5%-level. We therefore only report the standardized regression coefficients with 95% confidence intervals. Standardized regression coefficients can be interpreted like correlation coefficients and facilitate interpretation of how the two assessment features are comparatively related to either Cronbach's Alpha, as an estimate of reliability, or classification accuracy. Simulation results are publicly available for interested researchers (https://surveybanken.sikt.no/en/study/NSD3183).

RESULTS

We first present the descriptive statistics for the fifteen real datasets from actual exams. Then, we present the results of the 52,500 resampled conditions descriptively for each of the three exams included. As part of this, we provide more detailed comparison of the difference between estimates of classification accuracy and reliability. We present these graphically and report the results of the statistical modelling - for alpha and then for classification accuracy respectively. We end the results section with a summary of our key findings.

Descriptive statistics

Descriptive statistics for the fifteen original exams

The grand mean of the percentage correct score across the original fifteen was M = 71% (SD = 3%). Number of participating students was, on average N = 121 with range between min = 89 and max = 189. Number of items ranged between 79 and 143, with a median of 108

items per exam. Cronbach's Alpha was on average Alpha = .78, ranging between Alpha(min) = .63 and Alpha(max) = .92. For the original exams, the average fail rate was 3.83% of the candidates, ranging from zero percent to a maximum of 13.1% on one single occasion. This occasion was, notably, the first regular exam after a series of remote exams due to COVID-19-related restrictions.

Descriptive statistics for the resampled exams

Mean classification accuracy across all conditions was $M_{CA} = 0.93$ (SD = 0.06) for the non-parametric approach meaning that 93% of pass/fail decisions were estimated as accurate across all data. The average reliability was Alpha = .73 (SD = 0.16) implying in one interpretation that across all data the scores on the tests correlate 0.73 with a 'perfect' test. The two measures—reliability and classification accuracy—were positively associated, with a correlation of $r = 0.28$ ($t = 66$, $df = 52498$, $p < .001$). This suggests that these two indices are only moderately aligned—they provide different information about the psychometric properties of the exam.

Descriptive statistics and inter-estimate associations for all five coefficients are presented in Table 2. These results show that, across all conditions, the three different measures of classification accuracy are closely aligned, with the lowest correlation coefficient being $r = 0.95$ ($t = 966$, $df = 52490$, $p < .001$) for non-parametric classification accuracy and the CTT-based approach (i.e., Livingston & Lewis). The same pattern of high inter-estimate correlations is found for the two reliability coefficients - across all 52,500 resampled exams, Cronbach's Alpha and IRT-based EAP reliability correlated $r = 0.96$ ($t = 781$, $df = 52498$, $p < .001$). This suggest that the choice of an overarching psychometric framework (CTT or IRT) makes little difference to our substantive conclusions.

Relation between cut score, test length and Cronbach's Alpha

For the relation of Cronbach's Alpha and test length, we found an effect of $\beta_{std} = 0.81$ (95% CI [0.80 , 0.81]; cf. Table 3) using the linear mixed effects model. As expected, the cut score as a predictor was not related to variation in Cronbach's Alpha ($\beta_{std} = 0.00$, 95% CI [-0.01 , 0.00]).

Detailed results are given in Figure 1 which shows how classification accuracy (lower panel) and Cronbach's Alpha (upper panel) vary in relation to test length (number of items; horizontal axes) for a range of cut-scores (40%, 50%, 60%, 70% and 80%). On average, both estimates increase with test length. Regarding Cronbach's Alpha, the upper panel of Figure 1 illustrates visually how Alpha is not affected by the various cut-scores. At the same time, Alpha varies mainly due to the number of items in the resampled exams. Furthermore, the intra class correlation (ICC) for the exam-level effect was 0.19 which indicates that 19% of the variation in Cronbach's Alpha is due to this clustering in exams. Put differently, within the three exams, estimates of Cronbach's Alpha tend to be more similar. The model overall explained around 70% of the total variation in the estimate.

Since visual inspection of the descriptive results (i.e., Figure 1) suggested a non-linear trend for the effect of test length on Cronbach's Alpha, we also estimated an additional model in which we included both a linear ($\beta_{std} = 0.89$, 95% CI [0.88 , 0.89]) and a quadratic trend ($\beta_{std} = -0.38$, 95% CI [-0.38 , -0.37]) for test length. This model increased the variance explained in Cronbach's Alpha to a total of 79.3%.

Relation between cut score, test length and classification accuracy

For classification accuracy, when we only entered test length in the regression model, we found a standardized regression coefficient of $\beta_{std} = 0.31$ (95% CI [0.30 , - 0.31]) for the relation between test length and classification accuracy (Table 3). This is a weaker effect than that for Cronbach's alpha ($\beta_{std} = 0.81$ cf. paragraph above). Generally, higher cut-scores were related to lower classification accuracy ($\beta_{std} = -0.79$, 95% CI [- 0.77 , - 0.75], cf. Table 4). Since a higher-cut score moves closer to the mean of the score distribution, a more equal balance of pass-fail decisions is made, which in turn implies a greater likelihood of incorrect classifications. This pattern is shown in Figure 1, lower panel, where an exam with 20 items, where most students pass (1% fails), has a higher classification accuracy than a 150-item-exam where about half of students pass (46% fails).

The ICC for the exam-level effect was low with ICC = 0.02; indicating that the exam-level variation, that is, similarity in estimates which are due to the clustering within the three exams, was negligible, and a weaker effect than seen for Cronbach's alpha. In total, the model explained almost 62% of the variation in the non-parametric estimate for classification accuracy. For completeness and to be consistent with the reliability analysis, we also adding a quadratic trend for the test length. The variance explained by the model increased by less than one percentage point.

Summary of key findings and differences in test length guidelines

In general, our results indicate that reliability and classification accuracy are differently related to test length and that our initial hypothesis that test length decisions based on classification accuracy would generally be shorter is too simplistic. However, as expected, only classification accuracy varies in relation to the cut-score for pass-fail decisions. Based on our analyses, to optimize the reliability of a particular exam, we find by inspection that optimal test

length, across conditions, is around 100 items. At this test length, estimates are close to or exceed the 'standard recommendation' of Alpha = .8 (Figure 1, upper panel). For classification accuracy, however, the derived recommendations are less generic (Figure 1, lower panel). For exams with a small percentage of fail decisions (i.e., 5% or less, typical for many national licensing exams), an item size of 50 would typically achieve an accuracy of 95% correct classifications. With an increase in the cut-score and, accordingly, a higher share fail-decisions, much longer exams would be recommended. For instance, for 18% fail-decisions, a 150-item exam would reach, on average, a level of approx. 90% accuracy. Importantly, these results indicate that given the purpose of the exam ('ranking' vs. 'competent or nor', which are mirrored in the psychometric indicator used), we reach different conclusions regarding an optimal test design.

DISCUSSION

Our study is based on the notion that different perspectives on the precise purpose of an assessment can lead to different recommendations for the test length of assessments. In addition, we comment that there seem to be two qualitatively different strands in the literature. One is psychometrics-focused and essentially bases recommendations for test lengths on reliability calculations. The other strand is more validity-focused and put emphasis on 'broad sampling' of domains [21]. Only the former approach has proposed clear recommendations for the specific number of items (or stations) needed in an exam (see for example, van der Vleuten and Schuwirth[7]). Against this background, our work shows that for contexts where classification into competent or not competent are most important, shorter tests are often sufficiently accurate and

446 defensible. Crucially, this depends upon the cut-score relative to the average candidate
447 performance, or equivalently, the expected failure rate.

448 The results provided here have practical implications for the design of both individual
449 exams and programmes of assessment, both within HPE institutions, and for national licensing
450 examinations[6,27,28]. In scenarios where only a low proportion of failing students is expected, our
451 work suggests that shorter exams might reach sufficiently high levels of classification accuracy
452 for high stakes decision-making. Opportunities to re-design assessments might be therefore most
453 promising where exams tend to have lower failure rates. For instance, less comprehensive testing
454 might be needed at the end of medical training or at the end of larger modules of teaching.
455 However, our results indicate that test design decisions guided by reliability calculations are
456 generally likely to require more items in such contexts. Importantly, our work suggests that the
457 conclusions on test length drawn from reliability estimates are usually in the opposite direction
458 of those from classification accuracy.

459 Taken together, based on our analysis, we can make the following specific
460 recommendations for test developers involved in HPE.

461 • If you mainly make pass-fail decisions in your exams then, in addition to
462 reliability, use one of the free tools to also calculate classification accuracy (more
463 on this below).
464 • Think about how many items are appropriate to achieve sufficient coverage of the
465 assessed domain. Is the number of items justified as achieving appropriate levels
466 of reliability or to secure appropriate construct representation?
467 • Where possible, reduce the assessment burden on item writers by shortening the
468 exam. For instance, in a context similar to the one here, consider reducing the

number of items from 100 to 80. This will allow for a focus on better item quality rather than greater volume.

Clearly, test design involves many consequential decisions. In our opinion, these should not be solely based on psychometric indicators since influential thinking on validity stresses the need for appropriate representation of the assessed constructs [16]. If, for instance, one hundred items are generally considered the minimum necessary for appropriate representation in a particular area of HPE, adjusting that number downwards would not be justified. However, appropriate representation is both a matter of the choice of the tasks included ('which'/'how much') and their fidelity ('how realistic') [16]. In practice, these demands might clash where there are limited resources for producing high quality assessment content. Even where item pools are available and large in size, maintaining the quality of the individual items might be problematic, particularly in rapidly changing areas of medicine for example.

Another practical implication of our research is that more innovative testing models and assessment designs might be facilitated by a shift towards, or at least greater emphasis on, classification accuracy over reliability. For instance, this line of thinking would support more frequent, but spaced exams, instead of relying on massed 'big' exams, thereby capitalizing on spaced learning effects [53]. Importantly, this could be done without the need for large item banks, advanced psychometric analyses, or the more technical demands of computerized adaptive testing. Ultimately, such developments would likely enable better learning for students [54].

On a theoretical level, our findings suggest a need to estimate and discuss appropriate levels of classification errors in decision-making in assessment. By the nature of assessment and testing, such errors are nearly always made, with some students who are not ready moving on to the next step in training, and competent students failing exams. At the very least, classification

accuracy estimates should be presented and discussed among the decision makers involved in overseeing the assessments. For those unfamiliar with calculating classification accuracy, we suggest using a web application based on one of the R packages used here[47]. This is an open source online-tool and straightforward to use with raw assessment scores [https://hthaa.shinyapps.io/shinybeta/].

Discussions on the use of appropriate psychometric models in the assessment of medical competence have often focused on concepts such as latent traits or reliability indices [1,8,55]. Our study focusses on just one other—classification accuracy—of many quantitative, psychometric concepts that can be inform the design of assessment [56]. However, in this work we have only really touched on the topic of classification accuracy , while assessment policy-making would also need to consider more carefully the balance of likely decision errors (e.g., false positives or negatives) or other assessment properties such as its sensitivity and specificity [57]. More detailed classification accuracy analysis than presented here can produce useful metrics for these properties. In a clinical, rather than assessment, context, these technical measures are typically evaluated against a gold standard (e.g., results of another diagnostic test), but in educational settings this gold standard is typically not available. Fortunately, psychometric theory allows us to estimate these different indices based on certain statistical assumptions.

As our results indicate, strong correlations across different estimates of either reliability or classification accuracy suggest that the choice of the underlying measurement framework (CTT or IRT) matters little. Rather, what matters most is that we explicitly consider the alignment of the purpose of an exam, the use of the scores obtained, and the psychometric concepts used. Importantly, our study illustrates that the proper use of psychometric methods is not self-evident. This echoes the strong stances put forward by leading scholars in the field.

Recently, Robert Brennan cautioned against a simplistic interpretation of reliability stating that "[c]learly, reliability is a ratio that depends on both the modeling of observed scores and the definition of error, which renders the concept of reliability to be far more challenging than typically understood. Indeed, Cronbach (2004), Kane (1996), and this author advise abandoning routine use of reliability coefficients because they are so easily misinterpreted."[58]

An obvious limitation of this work is that it is based on the resampling of knowledge test assessment data from a single institution, rather than real test data from a range of institutions with varying test lengths. Whilst we have made attempts to make the best use of this data via our resampling approach, future work could aim to derive practical recommendations for classification accuracy guidelines on a broader selection of different types of exam data, in particular OSCE-type assessments as well as combinations of observations as in programmatic assessments. Consequently, future research could make use of assessment data from a broader range of institutions. It is likely that data that includes information on the sequence and time a student used for items in an exam gathered automatically in an online platform would be highly informative. This data could be employed to get a more authentic estimate of the by-item increase in reliability and classification accuracy in the course of testing, and might provide an additional perspective that does not suffer from the 'artificiality' of our resampling procedure. For more complex assessments such as the OSCE or a set of mini-CEX encounters, more complex psychometric procedures to calculate classification accuracy are available[36,59] and could be compared to estimates of reliability just as in the case provided here. In these more complex contexts, the issue of domain representation is likely to be more acute than in an applied knowledge test.

The evidence suggests that when it comes to assessment design issues, test developers and assessment policy individuals need to think carefully and explicitly about the precise purpose(s) of the assessment. Shifting our perspective on assessment towards accurate classification also requires being explicit about our expectations about student cohort groups, about base-rates of students being competent (or not), and how this might change as training proceeds. This shift in perspective also suggests that reliability considerations alone are likely too simplistic, and that high-reliability exams might sometimes lack appropriate levels of classification accuracy. Unfortunately, our work suggests that there are no very simple guidelines on test length and appropriate indices that can be stated to cover all cases – possibly a message from our work that busy faculty are not going to find of comfort. As a first step, we would advocate for the use of both reliability and classification indices when 'assessing the assessment' [60], with the inferences made about the quality of the test and associated outcomes also depending on careful consideration of the particular context.

552

553 **REFERENCES**

554

555 1. Schuwirth LWT, van der Vleuten CPM. A history of assessment in medical education. *Adv*
556 *Health Sci Educ*. 2020;25(5):1045-1056. doi:10.1007/s10459-020-10003-0

557 2. Case SM, Holtzman K, Ripkey DR. Developing an Item Pool for CBT: A Practical
558 Comparison of Three Models of Item Writing. *Acad Med*. 2001;76(10):S111.

559 3. Karthikeyan S, O'Connor E, Hu W. Barriers and facilitators to writing quality items for
560 medical school assessments – a scoping review. *BMC Med Educ*. 2019;19(1):123.
561 doi:10.1186/s12909-019-1544-8

562 4. Moonen-van Loon JMW, Overeem K, Donkers HHLM, van der Vleuten CPM, Driessen
563 EW. Composite reliability of a workplace-based assessment toolbox for postgraduate
564 medical education. *Adv Health Sci Educ*. 2013;18(5):1087-1102. doi:10.1007/s10459-013-
565 9450-z

566 5. Swanson DB, Norman GR, Linn RL. Performance-Based Assessment: Lessons From the
567 Health Professions. *Educ Res*. 1995;24(5):5-11. doi:10.3102/0013189X024005005

568 6. Swanson DB, Roberts TE. Trends in national licensing examinations in medicine. *Med Educ*.
569 2016;50(1):101-114. doi:10.1111/medu.12810

570 7. Van Der Vleuten CPM, Schuwirth LWT. Assessing professional competence: from methods
571 to programmes. *Med Educ*. 2005;39(3):309-317. doi:10.1111/j.1365-2929.2005.02094.x

572 8. Pearce J, Tavares W. A philosophical history of programmatic assessment: tracing shifting
573 configurations. *Adv Health Sci Educ*. 2021;26(4):1291-1310. doi:10.1007/s10459-021-
574 10050-1

575 9. Pearce J, Chiavaroli N, Tavares W. On the use and abuse of metaphors in assessment. *Adv*
576 *Health Sci Educ*. Published online February 2, 2023. doi:10.1007/s10459-022-10203-w

577 10. Ryan MS, Holmboe ES, Chandra S. Competency-Based Medical Education: Considering Its
578 Past, Present, and a Post–COVID-19 Era. *Acad Med*. 2022;97(3S):S90.
579 doi:10.1097/ACM.0000000000004535

580 11. Livingston SA, Lewis C. Estimating the Consistency and Accuracy of Classifications Based
581 on Test Scores. *J Educ Meas*. 1995;32(2):179-197. doi:10.1111/j.1745-3984.1995.tb00462.x

12. Rudner LM. Expected Classification Accuracy. *Pract Assess Res Eval*. 2005;10(1). doi:10.7275/56a5-6b14

13. Norcini J, Anderson B, Bollela V, et al. Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 Conference. *Med Teach*. 2011;33(3):206-214. doi:10.3109/0142159X.2011.551559

14. Norcini J, Anderson MB, Bollela V, et al. 2018 Consensus framework for good assessment. *Med Teach*. 2018;40(11):1102-1109. doi:10.1080/0142159X.2018.1500016

15. American Educational Research Association. *Standards for Educational and Psychological Testing*. American Educational Research Association; 2014.

16. Messick S. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *Am Psychol*. 1995;50(9):741-749. doi:10.1037/0003-066X.50.9.741

17. Downing SM. Threats to the Validity of Locally Developed Multiple-Choice Tests in Medical Education: Construct-Irrelevant Variance and Construct Underrepresentation. *Adv Health Sci Educ*. 2002;7(3):235-241. doi:10.1023/A:1021112514626

18. Tavakol M, Dennick R. Post-examination interpretation of objective test data: Monitoring and improving the quality of high-stakes examinations: AMEE Guide No. 66. *Med Teach*. 2012;34(3):e161-e175. doi:10.3109/0142159X.2012.651178

19. Raymond MR. A Practical Guide to Practice Analysis for Credentialing Examinations. *Educ Meas Issues Pract*. 2002;21(3):25-37. doi:10.1111/j.1745-3992.2002.tb00097.x

20. Raymond MR, Neustel S. Determining the Content of Credentialing Examinations. In: *Handbook of Test Development*. Lawrence Erlbaum Associates Publishers; 2006:181-223.

21. Schuwirth L, van der Vleuten C. How to design a useful test: the principles of assessment. In: Swanwick T, Forrest K, O'Brien BC, eds. *Understanding Medical Education*. 3rd ed. Wiley-Blackwell; 2018:275-289. doi:10.1002/9781119373780.ch20

22. Park YS. Reliability. In: Yudkowsky R, Park YS, Downing SM, eds. *Assessment in Health Professions Education*. 2nd ed. Routledge; 2019:33-50. doi:10.4324/9781138054394

23. Downing SM. Reliability: on the reproducibility of assessment data. *Med Educ*. 2004;38(9):1006-1012. doi:10.1111/j.1365-2929.2004.01932.x

24. Karay Y, Schauber SK. A validity argument for progress testing: Examining the relation between growth trajectories obtained by progress tests and national licensing examinations using a latent growth curve approach. *Med Teach*. 2018;40(11):1123-1129.

25. Wrigley W, van der Vleuten CPM, Freeman A, Muijtjens A. A systemic framework for the progress test: strengths, constraints and issues: AMEE Guide No. 71. *Med Teach*. 2012;34(9):683-697. doi:10.3109/0142159X.2012.704437

26. West CP, Durning SJ, O'Brien BC, Coverdale JH, Roberts LW. The USMLE Step 1 Examination: Can Pass/Fail Make the Grade? *Acad Med*. 2020;95(9):1287-1289. doi:10.1097/ACM.0000000000003537

27. Gedamu Wonde S, Schauber SK. Psychometric properties of the Ethiopian national licensing exam in medicine: an analysis of multiple-choice questions using classical test theory. *Teach Learn Med*. 0(0):1-11. doi:10.1080/10401334.2024.2428191

28. Guttormsen S, Beyeler C, Bonvin R, et al. The new licencing examination for human medicine: from concept to implementation. *Swiss Med Wkly*. 2013;143(4950):w13897-w13897. doi:10.4414/smw.2013.13897

29. Gomboo A, Gombo B, Munkhgerel T, Nyamjav S, Badamdorj O. Item Analysis of Multiple Choice Questions in Medical Licensing Examination. *Cent Asian J Med Sci*. 2019;(2):141-148. doi:10.24079/CAJMS.2019.06.009

30. GMC. How we assess doctors new to UK practice is changing, here's why. Supporting good, safe patient care across the UK. April 27, 2023. Accessed April 16, 2025. https://gmcuk.wordpress.com/2023/04/27/how-we-assess-doctors-new-to-uk-practice-is-changing-heres-why/

31. Sijtsma K. On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha. *Psychometrika*. 2009;74(1):107-120. doi:10.1007/s11336-008-9101-0

32. Lathrop QN, Cheng Y. Two Approaches to Estimation of Classification Accuracy Rate Under Item Response Theory. *Appl Psychol Meas*. 2013;37(3):226-241. doi:10.1177/0146621612471888

33. Lathrop QN, Cheng Y. A Nonparametric Approach to Estimate Classification Accuracy and Consistency. *J Educ Meas*. 2014;51(3):318-334. doi:10.1111/jedm.12048

34. Schauber SK, Hecht M. How sure can we be that a student really failed? On the measurement precision of individual pass-fail decisions from the perspective of Item Response Theory. *Med Teach*. 2020;42(12):1374-1384. doi:10.1080/0142159X.2020.1811844

35. Downing SM. Validity: on the meaningful interpretation of assessment data. *Med Educ*. 2003;37(9):830-837. doi:10.1046/j.1365-2923.2003.01594.x

36. Setzer JC, Cheng Y, Liu C. Classification Accuracy and Consistency of Compensatory Composite Test Scores. *J Educ Meas*. 2023;60(3):501-519. doi:10.1111/jedm.12357

37. Schauber S, Stensløkken KO. No knowledge gap in human physiology after remote teaching for second year medical students throughout the Covid-19 pandemic. *No Knowl Gap Hum Physiol Remote Teach Second Year Med Stud Covid-19 Pandemic*. 2023;23(976). http://www.biomedcentral.com/bmcmededuc/

38. University of Oslo. How to classify data and information. July 6, 2024. https://www.uio.no/english/services/it/security/lsis/data-classes.html

39. Wennberg A. Kan eksamen bli kortere og bedre? https://www.med.uio.no/imb/forskning/aktuelt/aktuelle-saker/2024/bedre-og-kortere-eksamen.html

40. Schauber SK. Re-Designing Assessments—Recommended Test-Length Based on Estimates of Either Reliability or Classification Accuracy (Version 1). https://doi.org/10.18712/NSD-NSD3183-V1

41. James G, Witten D, Hastie T, Tibshirani R, Taylor J. Resampling Methods. In: James G, Witten D, Hastie T, Tibshirani R, Taylor J, eds. *An Introduction to Statistical Learning: With Applications in Python*. Springer International Publishing; 2023:201-228. doi:10.1007/978-3-031-38747-0_5

42. Aubin AS, Young M, Eva K, St-Onge C. Examinee Cohort Size and Item Analysis Guidelines for Health Professions Education Programs: A Monte Carlo Simulation Study. *Acad Med*. 2020;95(1):151. doi:10.1097/ACM.0000000000002888

43. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; 2023. https://www.R-project.org/

44. Raju NS, Price LR, Oshima TC, Nering ML. Standardized Conditional SEM: A Case for Conditional Reliability. *Appl Psychol Meas*. 2007;31(3):169-180. doi:10.1177/0146621606291569

45. Robitzsch A, Kiefer T, Wu M. *TAM: Test Analysis Modules*.; 2022. https://CRAN.R-project.org/package=TAM

46. Lathrop QN. *cacIRT: Classification Accuracy and Consistency under Item Response Theory*.; 2015. https://CRAN.R-project.org/package=cacIRT

47. Haakstad HE. *Betafunctions: Functions for Working with Two- And Four-Parameter Beta Probability Distributions and Psychometric Analysis of Classifications*.; 2022. https://CRAN.R-project.org/package=betafunctions

48. Kane M. The Precision of Measurements. *Appl Meas Educ*. 1996;9(4):355-379. doi:10.1207/s15324818ame0904_4

49. Kane M. Using Error/Tolerance Analysis to Design an Empirical Practice Analysis. *Adv Health Sci Educ*. 2000;5(3):179-196. doi:10.1023/A:1009821413152

50. Kane M. The Errors of Our Ways. *J Educ Meas*. 2011;48(1):12-30. doi:10.1111/j.1745-3984.2010.00128.x

51. Brennan RL. Raw-score conditional standard errors of measurement in generalizability theory. *Appl Psychol Meas*. 1998;22(4):307-331. doi:10.1177/014662169802200401

687 52. Bates D, Mächler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models Using lme4.
688 *J Stat Softw*. 2015;67(1):1-48. doi:10.18637/jss.v067.i01

689 53. Versteeg M, Hendriks RA, Thomas A, Ommering BWC, Steendijk P. Conceptualising
690 spaced learning in health professions education: A scoping review. *Med Educ*.
691 2020;54(3):205-216. doi:10.1111/medu.14025

692 54. Lambers A, Talia AJ. Spaced Repetition Learning as a Tool for Orthopedic Surgical
693 Education: A Prospective Cohort Study on a Training Examination. *J Surg Educ*.
694 2021;78(1):134-139. doi:10.1016/j.jsurg.2020.07.002

695 55. Schauber SK, Hecht M, Nouns ZM. Why assessment in medical education needs a solid
696 foundation in modern test theory. *Adv Health Sci Educ*. 2018;23(1):217-232.

697 56. Collares CF. Cognitive diagnostic modelling in healthcare professions education: an eye-
698 opener. *Adv Health Sci Educ*. 2022;27(2):427-440. doi:10.1007/s10459-022-10093-y

699 57. Lalkhen AG, McCluskey A. Clinical tests: sensitivity and specificity. *Contin Educ Anaesth*
700 *Crit Care Pain*. 2008;8(6):221-223. doi:10.1093/bjaceaccp/mkn041

701 58. Brennan RL. Current Psychometric Models and Some Uses of Technology in Educational
702 Testing. *Educ Meas Issues Pract*. 2024;43(4):88-92. doi:10.1111/emip.12644

703 59. Lee WC. Classification Consistency and Accuracy for Complex Assessments Using Item
704 Response Theory. *J Educ Meas*. 2010;47(1):1-17. doi:10.1111/j.1745-3984.2009.00096.x

705 60. Pell G, Fuller R, Homer M, Roberts T. How to measure the quality of the OSCE: A review
706 of metrics – AMEE guide no. 49. *Med Teach*. 2010;32(10):802-811.
707 doi:10.3109/0142159X.2010.507716

708

709

FIGURES & TABLES

Figure 1 *** ENTER AS SIDEWAYS FIGURE ***

Relationship between number of items and reliability (Cronbach's Alpha) as a function of the cut-score.

[ENTER Figure 1.pdf HERE] *** ENTER AS SIDEWAYS FIGURE ***

Note. Lower panel shows relationship between classification accuracy and number of items as a function of the cut-score. Bold, black dots and lines are the grand mean across all resampled exams in each condition. The average across 100 resamples of one specific condition of exam, cut-score, and item-size is marked by an X. Classification Accuracy is the non-parametric approach (Lathrop & Cheng, 2014).

Table 1

Pseudo-code for the resampling procedure that was applied to each of the fifteen exams

```
for nitems in [20, 30, 50, 75, 100, 125, 150]
     →  for cutscore in [40, 50, 60, 70, 80]

          →  for i in [1 to 100]

               →  draw a random sample DATA[i] of nitems

                    →  CALCULATE ACCURACY for cutscore in DATA[i]

                    →  CALCULATE ALPHA for DATA[i]
```

Note. Pseudo-code for the resampling procedure that was applied to each of the fifteen exams, which indicates a loop across different conditions in our procedure: nitems can take the number of 20, 30… up to 150. Similarly, cutscore accounts for the varying contitions between 40 and 80. The running index i indicates that for each combination of number of items [nitems] and pass mark (cutscore), one hundred random samples are drawn. And for each of these samples, both classification accuracy (CA) and Cronbach's Alpha (ALPHA) are calculated.

Table 2

Means, standard deviations, and correlations with confidence intervals for the estimates of reliability and classification accuracy across all 52,500 resampled exams.

Coefficient	M	SD	1	2	3	4
1. Cronbach's Alpha	0.73	0.16				
2. IRT-based Reliability	0.72	0.18	.96** [.96, .96]			
3. Non-Parametric classification accuracy	0.93	0.06	.28** [.27, .28]	.25** [.24, .26]		
4. L&L classification accuracy	0.92	0.07	.32** [.31, .33]	.29** [.28, .30]	.95** [.95, .95]	
5. Rudner classification accuracy	0.93	0.06	.36** [.35, .37]	.33** [.33, .34]	.97** [.96, .97]	.96** [.96, .96]

Note. M and SD are used to represent mean and standard deviation, respectively. Values in square brackets indicate the 95% confidence interval for each correlation. * indicates $p < .05$. ** indicates $p < .01$.

Table 3

Linear mixed effects models which include assessment features as predictors for Cronbach's Alpha or Classification Accuracy

	Cronbach's Alpha	**Classification Accuracy**
	β [CI]	β [CI]
(Intercept)	-0.00 [-0.30 – 0.30]	-0.00 [-0.09 – 0.09]
Test length	0.81 [0.80 – 0.81]	0.31 [0.30 – 0.31]
Cutscore	-0.00 [-0.01 – 0.00]	-0.72 [-0.72 – -0.71]
Random Effects		
Exam-level variance τ_{00}	17.16 $_{exam}$	0.22 $_{exam}$
Residual variance σ^2	73.19	13.31
Exam-level ICC	0.19	0.02
Marginal R^2 / Conditional R^2	0.639 / 0.708	0.611 / 0.618

Note. β is the standardized regression coefficient. CI is the 95% confidence interval. σ^2 is residual variance. τ_{00} is the random effect for the exam-factor. ICC is the intra-class correlation. Marginal R^2 is the variance explained by the fixed effects only. Conditional R^2 is the variance explained by both random and fixed effects.

Disclosures

Acknowledgments

Both authors made substantial contributions to the conception of the study and design of the analyses as well as to the interpretation of the data. SKS was responsible for data acquisition and ran the analysis. SKS provided the first draft of the study and MH provided critical revisions and contributed with important intellectual content. Both authors gave approval for submitting the current work and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Funding/Support

None

Other disclosures

None

Ethical approval

The Norwegian Agency for Shared Services in Education and Research approved the processing of archival exam data on the grounds of research in public interest (General Data Protection Regulation art. 6 nr .1 e). Reference number 497365.

Disclaimers

None

Previous presentations

None

Data

None