Evaluating the Quality of Higher Education Instructor-constructed Multiple-Choice Tests: Impact on Student Grades

Gavin T L Brown

The University of Auckland

Hasan H. A. Abdulnabi

AMA International University (Bahrain)

Correspondence concerning this paper can be addressed to Professor Gavin T L Brown, The Faculty of Education and Social Work, The University of Auckland, Private Bag 92019, Auckland 1142 New Zealand or by email to gt.brown@auckland.ac.nz

Evaluating the Quality of Higher Education Instructor-constructed Multiple-Choice Tests: Impact on Student Grades

ABSTRACT

Multiple-choice questions (MCQs) are commonly used in higher education assessment tasks because they can be easily and accurately scored, while giving good coverage of instructional content in a short time. However, studies that have evaluated the quality of MCQs used in higher education assessments have found many flawed items, resulting in misleading insights about student performance and contaminating important decisions. Thus, MCQs need to be evaluated statistically to ensure high quality items are used as the basis of inferences. This study evaluated the quality of 100 instructor-written MCQs used in an undergraduate midterm test (50 items) and final exam (50 items), making up 50% of the course grade, using the responses of 380 students enrolled in one 1st-year undergraduate general education course. Item difficulty, discrimination, and chance properties were determined using Classical Test Theory and Item Response Theory statistical item analysis models. The two-parameter logistic model consistently had the best fit to the data. The impact on overall course grades between the original raw score model and the IRT 2PL model showed 70% of students would receive the same grade (i.e., D to A), but only one-third would get the same mark using the standard augmented grade scale (i.e., A+ to D-). The analyses show that higher education institutions need to ensure MCQs are evaluated before student grading decisions are made.

Multiple-choice questions (MCQs) are one of the most commonly used assessment methods in higher education (Bailey, Mossey, Moroso, Cloutier, & Love, 2012; DiBattista & Kurzawa, 2011). However, use varies by discipline (e.g., high use is seen in medical education) and jurisdiction (e.g., perhaps less so in the United Kingdom). MCQs are used because, in a short period of time, a broad range of course material can be efficiently assessed and accurately scored (DiBattista & Kurzawa, 2011; Nedeau-Cayo, Laughlin, Rus, & Hall, 2013). However, the quality of the instructor-written MCQs used in higher education assessments is questionable and potentially results in misleading evidence of student achievement (Brady, 2005; Downing, 2005; Masters et al., 2001; Stagnaro-Green & Downing, 2006; Tarrant, Knierim, Hayes & Ware, 2006). This is an understandable situation since few academics have had formal education in assessment theory or the principles of MCQ item writing.

Evaluation of MCQ quality can be conducted through professional judgment processes relative to 'best practice' conventions and advice (Haladyna, 2004). Four major foci have been identified: (1) content guidelines, (2) style and format, (3) writing the stem, and (4) writing options (Haladyna & Rodriguez, 2013). Implementation of these guidelines can and should be conducted automatically 'in-house' by academics within each department or discipline, prior to deployment of the test or examination. The second approach for determining quality is the application of statistical item analysis procedures to determine the characteristics of items and use those statistics to decide if an item can be properly included in the determination of test-taker performance (Downing, 2006; Malau-Aduli, & Zimitat, 2012).

Items which are (a) inappropriately difficult or easy, (b) too easy to guess at, or (c) do not discriminate positively between high and low performing learners will lead to inappropriate decisions about student ability and consequent decisions (e.g., pass-fail, graduation, access to scholarship, etc.). Furthermore, such items will also give inappropriate feedback to students and instructors. In both cases, the problem lies in poorly constructed items, rather than necessarily poorly delivered teaching or poor learning habits and strategies.

Statistical tools exist to evaluate item quality and are used extensively in high-stakes testing programs in international K-12 test systems, in national K-12 testing programs, and in high-stakes university admission testing. Unfortunately, the same cannot be said for higher education course assessments, especially those relying on MCQs. This is problematic since grades are awarded, in part, on the basis of performance on MCQ testing and, if little or no quality assurance is carried out, then invalid conclusions about student performance will be drawn. Hence, the lack of quality assurance processes, such as statistical item analysis or item evaluation, raises doubt as to the validity and legitimacy of scores, grades, and ultimately certificates and degrees. Thus, the goal of this study was to examine two operational MCQ-based tests within one course, using multiple statistical models, to determine (1) the quality of items and (2) possible implications for grading decisions. It is worth noting that, while this study involves MCQ items, the same issues and challenges exist for any dichotomously scored test question formats such as True-False, Mix-and-Match, and so on. The statistical problems are very similar, although the item quality indicators would be different.

Assessment in Higher Education

Assessments in higher education serve a wide range of functions, including formative (e.g., how and what to improve on) and summative (e.g., pass-fail decisions, entry to restricted programs, scholarships, graduation, etc.) (Schaughency, Smith, Meer, & Berg, 2012; Yorke, 2009). Tested performance using MCQs is normally transformed into grades (e.g., A to D or E) which are meant represent the quality of students' performance and level of achievement (Yorke, 2009). Grades are signals of achievement and show students their areas of strength and weakness, and can inform instructors about the success of their teaching

(Brown, 2010; Joughin, 2009; Walvoord & Anderson, 2011; Yorke, 2009). Obviously, the quality of assessment matters so that inferences and decisions by students and instructors, as well as external stakeholders (e.g., employers), can be made on a robust basis (Grainger, Purnell, & Zipf, 2008).

Quality of MCQs in Higher Education

Despite the existence of guidelines for writing MCQs (Brady, 2005; Burton, 2005; Downing & Yudkowsky, 2009; Haladyna, 2004), studies have found many bad items and violations of recommended guidelines. Tarrant, Knierim, Hayes, and Ware (2006) evaluated 2770 MCQs used over a five-year period from 2001 to 2005 and concluded that nearly half (46%) of the items were bad because they violated item-writing guidelines. Similar outcomes in higher education assessments are reported across different disciplines (Downing, 2005; Ellsworth, Dunnell, & Duell 1990; Hansen & Dexter, 1997; Masters et al., 2001). Poorly written MCQs can negatively impact students' performance and achievement (Clifton & Schriner, 2010; Downing, 2005; Tarrant, Knierim, Hayes, & Ware, 2006).

There is also concern that MCQs do not assess higher-order thinking and focus too much on recall of knowledge (Downing, 2005; Malau-Aduli & Zimitat, 2012; Popham, 2011; Tarrant, Knierim, Hayes, & Ware, 2006; Walsh & Seldomridge, 2006). Additionally, creating good MCQs is time-consuming and it is particularly difficult to create good distractors, especially for higher order thinking objectives (Clifton & Schriner, 2010; Fellenz, 2004). It has been proposed that poor item writing, rather than an inherent characteristic of MCQs, accounts for their tendency to assess lower-order cognitive skills (Downing, 2005; Downing & Yudowsky, 2009; Malau-Aduli & Zimitat, 2012). Fortunately, training in MCQ item writing has produced significantly higher quality MCQs (Jozefowicz et al., 2002).

Since MCQs contribute to course grades, a high score on an easy test may artificially inflate student grades. Likewise, the reverse occurs if the test was overly difficult resulting in artificially depressed grades. The problem of ensuring that item and test difficulty aligns with appropriate standards is complicated especially if item difficulty is caused by poor writing. Without quality assurance and standard setting processes that take into account the difficulty of the test relative to the grade criteria, raw scores on a test have little meaning. Hence, MCQ tests and examinations need to be evaluated for the quality of the item writing and the statistical properties of the contributing items. Then, standards need to be derived for each test, using one of many methods available (Cizek, 2001), which map the test scores onto grade descriptors. While item quality and standard setting are complex human processes, the analysis of item properties is a more technically demanding statistical process.

Statistical Approaches to MCQ Item Quality

Two major classes of statistical methods can be applied to MCQs. These are known as classical test theory and item response theory. The former examines tests as entities, while the latter evaluates items in and of themselves.

Classical Test Theory

The classical test theory (CTT) approach determines item characteristics from the available observed data (Reynolds, Livingston & Wilson, 2009). CTT assumes that the total number of items answered correctly indicates the examinee's level of ability or knowledge (de Ayala, 2009; Schaughency, Smith, van der Meer, & Berg, 2012). In other words, students who get a higher proportion of items correct know more than those with a lower percentage correct. Most commonly, letter grades are associated with ranges of percentage correct (e.g., B=70% to 79%) or a pass-score can be set at a proportion correct (e.g., 60%).

CTT specifies that the score achieved by an individual examinee is equal to the sum of their theoretical true ability and the unobserved error component in the test. The proportion of candidates getting an item right (p) determines the difficulty of each item; items that are too easy (p>.80) or too hard (p<.20) are frequently rejected from a test as not providing useful information about candidate ability. Ideally, all items in a test discriminate positively between those who know most and those who know least. This is determined by examining the point-biserial correlation (r_{pb}), which is the correlation of the item to the total after the item has been removed from the total. In many testing situations, items which do not have a significantly positive value (r_{pb}>.20) are rejected, though any positive value indicates a small tendency for higher scoring candidates to get individual items correct more than the lower scoring candidates (Ebel & Frisbie, 1991).

Another quality indicator of MCQ is the efficiency of the wrong answer distractors. Distractors that get selected infrequently (e.g., <5% of test-takers choose it) are so implausible that they seem to attract only candidates randomly guessing (Haladyna & Downing, 1993). Options with low selection rates have been found up to nearly half of all items (Haladyna & Downing, 1993), between 30-40% of all items (Tarrant, Ware, & Mohammed, 2009), and as high as 75% (Hingorjo & Jaleel, 2012). Hence, identification of such options and their subsequent replacement or deletion could improve item quality.

Test quality is accepted, generally, if the estimate of reliability (e.g., Cronbach's alpha) is sufficient for the decisions being made. For example, for research purposes α>.70 is considered sufficient because the shared covariance of the items accounts for about 50% of the test score. However, in a high-stakes certification examination (e.g., Advanced Placement tests at the end of high school in the USA), very high reliability estimates (α >.90) are expected. Given the mean, standard deviation, and reliability estimate of a test, it is possible to calculate a standard error of measurement, which is the range of scores that each candidate would most likely get the next time they sat the exact same test (Harvill, 1991). The SEM indicates the number of marks a score could vary by chance, without any substantive change in the student's ability and should be used in making decisions about quality or change.

Thus, the CTT approach provides sufficient statistics to evaluate items for difficulty and discrimination.

However, in CTT examinee ability is ‘sample dependent’ meaning that if the test is hard the students will seem to be low achievers and vice versa (Hambleton, Swaminathan, & Rogers, 1993). Similarly, items have difficulty values totally dependent on the ability of the sampled test takers, and so a change in their ability will change the item difficulty. This means that items will have very different characteristics depending on who attempted them and what other questions were present.

Item Response Theory

Because a test is a sample of a domain of interest, the real focus of interest in assessment is the learner’s ability in the domain, independent of the set of items presented in a test. Hence, a modern class of statistics (i.e., item response theory [IRT]) has arisen which permits items to be given different difficulties, discrimination, and guessing characteristics independent of the test in which they are presented (Borsboom, 2005; Embretson & Reise, 2000; Hambleton, Swaminathan, & Rogers, 1993). IRT predicts the likelihood that an examinee with a specific ability level will correctly answer a specific item by defining the examinee’s ability in relation to the item characteristics (Embretson & Reise, 2000). This means that a person’s total score or ability can be estimated using a probabilistic formula based on the dual properties of the item and the test-taker’s performance (Hambleton & Jones, 1993).

All IRT models propose that the probability of answering an item has an S or ogive shape in which the probability of answering correctly increases probabilistically as the ability of the test-takers increases. The formula uses the natural log of the odds that an item is answered correctly over answered incorrectly. The S-shape of the item plot (i.e., probability of answering correctly on the vertical axis versus item difficulty and person ability on the horizontal axis) creates two asymptotes so that the probability approaches, but never reaches, certainty (*p*=1.00 versus *p*=0.00) even as ability reaches positive or negative infinity (Giblin, 1972). This shows that there is always the possibility for very low-ability students to answer an item correctly by chance, and vice versa. Since MCQs have multiple wrong answers, it is possible for a high-ability student to be misled and similarly, because the right answer is available, it is possible for a low-ability student to randomly select it.

The difficulty of an item is the point when the probability of answering the item correctly equals 50%. The ability of a person is defined as the difficulty of items for which the person has a probability of answering correctly at the 50% correct rate (Embretson & Reise, 2000). Unlike CTT, answering more questions correctly does not increase the overall ability estimate unless the items are hard. In other words, in IRT the person's score goes up, not by answering more a higher proportion of questions correctly, but by answering much harder questions. If the difficulty of items does not align well with the test takers' ability (e.g., too many easy or very hard questions relative to performance) then the accuracy of the estimated score decreases.

Within IRT there are three major models with increasing complexity (Embretson & Reise, 2000; Hambleton, Swaminathan, & Rogers, 1993; Osterlind & Wang, 2012).

One-parameter logistic model (1PL) or Rasch model. This approach assumes that all items have statistically equivalent discrimination and only differ in terms of their difficulty (Bond & Fox, 2007). The probability of guessing is assumed to be equal and very close to zero when ability is very low. Items are deemed to fit the Rasch model if the Chi-square (χ^2) index is statistically not significant (Bond & Fox, 2007). Only items which are statistically equivalent to each other in terms of their slope and lower asymptotes are retained by the model. This can inadvertently mean that items with very strong positive discrimination could be rejected, simply because they differ too much from the model (Houts, Edwards, Wirth, & Deal, 2016). It is important to note here that the Rasch model approach prioritises the model which requires all items to conform to the assumptions and be statistically equivalent to each other. This stands in contrast to the data-centric approach of IRT, which may use the same 1PL model as Rasch, but allows parameters to be freely estimated without constraining them to fit the a priori model. While some have argued that the imposition of the Rasch model assumptions is necessary to achieve ‘measurement’ (Bond & Fox, 2007), the data analysis in this study is neutral as to the philosophic assumptions associated with Rasch modeling.

The Rasch or 1PL model has been used to analyse the quality of multiple-choice items with mixed results. Some studies showed that the Rasch model fitted most items (Athanasou & Lamprianou, 2004) while others have found that the Rasch model did not fit most multiple-choice items (Divig, 1986; Leeson & Fletcher, 2003), most likely because of the overly restrictive requirement that all items have zero guessing and equal discrimination (Drasgow & Parsons, 1983; van de Vijver, 1986). Hence, there are doubts as to the sufficiency of the Rasch model for MCQ items where the possibility of guessing exists.

Two-parameter logistic model (2PL). The 2PL model includes item discrimination and item difficulty as factors that determine the item and test-taker characteristics (Embretson & Reise, 2000; Hambleton, Swaminathan, & Rogers, 1993; Thissen & Orlando, 2001). Items that have greater discriminatory power have steeper slopes at the 50% probability point. This means that only a small change in ability will produce a large change in probability of answering correctly. Highly discriminating items are useful, as in CTT, in differentiating between examinees of different ability, especially when distinctions relative to a cut-off score or grade boundary are required (Osterlind & Wang, 2012; Thissen & Orlando, 2001). The advantage of 2PL over Rasch is that items do not have to have equal discrimination rates.

Nonetheless, negative discrimination values, as in CTT, provide misleading information about the domain of interest, necessitating the removal of such items.

Like the 1PL model, the 2PL model does not account for the possibility of correctly answering the item by chance or guessing. The logic here is that if items are well-written, the probability of guessing should be less than the raw chance of randomly picking the right answer from a set of options. For example, a 4-option item could be answered correctly one in four times. If the chance value for such an item is actually 17%, then the effect of chance can be ignored. The 2PL model has been found to fit the data from well-designed MCQ reading comprehension items than the Rasch model (Leeson & Fletcher, 2003).

Three-parameter logistic model (3PL). The 3PL model extends IRT by including a parameter that represents the possibility of low ability examinees answering an item correctly due to chance (Crocker & Algina, 1986; Embretson & Reise, 2000; Hambleton, Swaminathan, & Rogers, 1993; Osterlind & Wang, 2012). MCQ items rarely have a lower asymptote at zero because even very weak students can answer items correctly by random processes. Assuming all items are completely independent of each other, the chance of getting any 4-option MCQ correct should be 25%. Unsurprisingly, if items are poorly written or if test-takers exercise very little effort (Wise & Smith, 2016), the probability of guessing correctly could be much higher. Thus, items which have chance values greater than the random rate of guessing are normally removed before score determination. Studies of MCQs with the 3PL model have found it to fit well most of items in a test (Adedoyin & Mokobi, 2013; Bergan, 2010; Leeson & Fletcher, 2003).

Sample Size. Given the complexity of the IRT models, it is not surprising that large sample sizes (i.e., $N \geq 500$) are considered necessary (Hambleton & Jones, 1993). This is in contrast to the smaller sizes (i.e., $N \geq 200$) permitted by the simpler approach of CTT. Claims that Rasch or 1PL models will estimate accurately with $N < 100$ (Boone, Staver, & Yale, 2014) have been found to be unreliable (Houts, Edwards, Wirth, & Deal, 2016). Research into real-data, as opposed to simulated data, with smaller sample sizes has suggested that $N<200$ is infeasible (Sireci, 1992), but that $N\geq300$ can provide reasonably accurate parameter estimation provided a test has ≥30 items (Akour & AL-Omari, 2013).

Thus, the challenge in operational testing of classes with teacher-made tests, where N<500 is commonplace, remains whether IRT techniques can be legitimately used to evaluate test items. In addition, few operational testing programs in higher education recycle items into future final examinations because institutional regulations require that exams be exposed to future students. This transparency practice means that once used, there is little opportunity to collect new data with the self-same items. Hence, in this naturalistic study, the accuracy of estimation given the available sample sizes has to be taken somewhat cautiously. Ideally, the results would be tested on a second independent sample for corroboration purposes, but this was beyond scope of the study.

Methods

Design

This study is a secondary data analysis of the course assessment data used to evaluate student learning in an introductory education course. The course was assessed with a mixture of two 50-item MCQ tests and essay examinations scored by course instructors and/or tutors. The course followed the university scoring system to convert percentage scores to grades against criterial descriptions for grades. Grades were A=excellent (80-100%); B=good (65-79%), C=satisfactory (50-64%), and D=unsatisfactory (0-49%). The minimum pass (C-) requires a score of at least 50%.

Participants

The participants in this study were students enrolled in one general education undergraduate course at a large research-intensive university in New Zealand. The course was an introductory educational psychology course on learning theory and offered as either a general education course (i.e., a course provided for students studying other disciplines outside the Faculty of Arts) or as a normal elective course (i.e., for students from the Faculty of Arts). General education courses in this study are first-year courses taken by students from outside the faculty hosting the course to broaden their education.

Of the 380 students for whom data were available, 276 were enrolled in the general education course, while 104 were enrolled as Arts normal elective students. Only 375 of 395 students did the mid-term test, and 377 took the final examination, resulting in 372 students who received a final grade. This size of sample is sufficient for CTT analysis and is close to the recommended threshold for IRT. Unfortunately, as is the case in real-world testing programs, it is not possible to administer the items again to increase the sample size. Hence, if the IRT results are poor quality, this may be partly attributable to the relatively low sample size.

No specific demographic characteristics of the students (i.e., their gender; age or ethnic group) were available. These operational data were released for secondary analysis without identifying information for the purpose of evaluating the test item writing.

Instruments.

The course consisted of three assessments (i.e., midterm-test worth 25%, essay assignment worth 25%, and final exam worth 50%).

Midterm test. The mid-term test was two-hours long and was administered in Week 6 of the course, covering material related to the course content presented in the first five two-hour lectures. The test had 50 four-option MCQs, drafted by the main course lecturer, and vetted for quality by the course coordinator and the faculty examination manager against accepted best-practice conventions for writing multiple-choice items. The midterm-test items covered content related to seven topics: that is, (1) cognitive processing (6 items), (2) forgetting (8 items), (3) general learning theory (5 items), (4) memory (8 items), (5) meta-cognition (10 items), (6) retrieval (7 items), and (7) schema (6 items). The test was administered on paper under invigilation.

Final exam. The final exam had 50 four-option MCQs constructed by three of the course lecturers, and were checked against best-practice recommendations by the course co-ordinator and faculty examination administrator. The final exam excluded topics covered in the midterm-test and covered five topics, each with 10 items, taught after the mid-term test. These were (1) motivation, (2) approaches to learning, (3) problem solving, (4) social structure, and (5) behaviourism or observational learning. The exam was administered on paper under invigilation. The MCQs were worth half of the total exam score (i.e., 25% of the total course score) with three essay questions making up the balance of marks.

Analyses

Analysis of a test presumes that items are unidimensional, which may not be the case when a test covers multiple topic areas. Dimensionality was checked with confirmatory factor analysis of a single factor with 50 items using the weighted least square estimator with robust standard errors and mean- and variance adjusted $\chi 2$ test statistic (WLSMV) was used (Finney & DiStefano, 2006) in Mplus version 7.4 (Muthén & Muthén, 1998-2015) to account for the dichotomous nature of the items was conducted. Current standards suggest that models do not need to be rejected if the root mean square error of approximation (RMSEA) is <.08, the weighted root-mean-square residual (WRMR) is close to 1.00, the comparative fit index (CFI) is >.90 (Fan & Sivo, 2007; Yu, 2002). Cronbach alpha estimates of internal reliability were conducted before and after removal of misfitting items to establish further evidence of total test score coherence. Psychometric analysis of the items was conducted using each statistical model. Once misfitting items were identified, student scores were recalculated, and the impact of on students' grades and pass/fail rates was explored.

IRT analyses. All item response theory analyses were conducted with the 'ltm' package in R which also reports model fit (Rizopoulos, 2006). This package produces a variety of comparable fit indices including log likelihood, BIC, and AIC values. Differences of AIC >10 indicate that the model with the smaller AIC has superior fit to the data; likewise, all models included by 95% confidence interval set (indicated by the sum of Akaike weights $\Sigma w i \geq .95$) are plausible equally well-fitting models (Burnham & Anderson, 2004).

Classical test theory (CTT). Items with *p*-values equal to either 1 or 0 (i.e., 1= 100% of the students answered the item correctly, 0 = 0% of the students answered the item correctly) were discarded. Items with point biserial correlation (r_{pb}) values below 0.19 were discarded. All values were determined through SPSS version 21.

IRT 1PL model. Each item's fit statistic values were found using the chi-square (χ^2) index of probability that the item data fit the Rasch model (Raykov & Marcoulides, 2006). Only items with statistically non-significant chi-square values (i.e., $p \geq .05$) were retained.

IRT 2PL model. Items with discrimination values lower than 0.19 were rejected.

IRT 3PL model. Because all MCQs had four options, items with chance >0.25 were rejected.

Grade effect. After removing misfitting items, each student's score was generated using the revised set of items. The scores were transformed from raw percentage (CTT) or logit value (IRT) to match the original raw score mean and standard deviation for the mid-term and examination separately. This was done because a standard setting exercise in which course lecturers set grade cut scores based on the revised set of items was not feasible. After transformation, the scores were added to the essay-based course and exam scores to generate a course total score. Based on this value, the number of students being awarded each grade (A to D) and pass-fail (A to C vs. D) was determined.

Results

Given that the tests covered multiple topics, each was examined for unidimensionality. Model fit for a single factor was mixed for the midterm test (χ^2=4213.62, *df*=1175, χ^2/*df*=3.586, *p*=.06; RMSEA=.083; CFI=.934; WRMR=2.394) and good for the final exam (χ^2=1309.05, *df*=1175, χ^2/*df*=1.114, *p*=.29; RMSEA=.017; CFI=.936; WRMR=0.982). This suggested that examining all items as contributors to a single latent ability measure was plausible. The mixed result for the midterm test suggests that inspection of item psychometric properties might reveal causes of relatively weak fit for unidimensionality.

Item Analysis

Table 1 shows the psychometric properties of the midterm test and final exam items according to the statistical model used. Highlighted values indicate items which failed to meet analytic standards for each approach.

In the midterm test, the CTT model rejected 28 items, the IRT 1PL model rejected 26 items rejected, the IRT 2PL rejected 24 items, and the IRT 3PL rejected just 20 items. It was noteworthy that in terms of distractor efficiency, 66% of items had all distractors with >5% selection, 15 items (30%) had only one distractor with a low selection rate, and just two (4%) had two bad distractors. In total, just one item (M45) was identified as misfitting by all four methods, indicating that the different approaches lead to quite different decisions about item quality. Model fit statistics indicated that the 2PL model had best fit (AIC=17653.32, Σw*i*=1.00) compared to the 3PL (AIC=19052.00) and Rasch (AIC=1962.39) models. The item characteristic curves for the IRT 2PL show that despite its superior fit to the data many items clearly have inverse discrimination slopes or very flat trajectories with very high intercepts at logit -4.00 (Figure 1).

<<insert Figure 1 about here>

In the final exam, many fewer items were identified as misfitting. The CTT model rejected 14 items for low discrimination and distractor efficiency indicated 70% of items had no low selection options, 12 items (24%) had one low selection distractors, and three items (6%) had two low selection options. The IRT 1PL model rejected 19 items, IRT 2PL just three item, and 19 by the IRT 3PL. Two items (E36 and E44) were rejected by all four methods. This suggests that the quality of items written for the final exam was probably better than that for the midterm test or else that the alignment of the items to the student ability was greater. The IRT 3PL retained the most items in the midterm test, while the 2PL kept the most in the final exam. However, as per the midterm test, the IRT 2PL had the best fit (AIC=18546.93, Σw*i*=1.00) compared to the 3PL (AIC=18578.89) and Rasch (AIC=18889.88) models. The item characteristic curves for the IRT 2PL show that, in accordance with its superior fit to the data, few items have inverse discrimination slopes or very high intercepts at logit -4.00 (Figure 2).

<<insert Figure 2 about here>

Therefore, it seems that using the 2PL model as the basis of analysing these two MCQ tests is the most robust approach and the Rasch method is the least effective IRT method. Nevertheless, in the context of classroom assessment, there may be a legitimate goal in including very easy or very difficult items (e.g., motivating sense of learning or establishing learning needs). These very easy or difficult items might have poor item discrimination statistics, but may still be useful to ensure an adequate sample of the constructs of interest.

<insert Table 1 about here>

Test Analysis

After removing the misfitting items, the test statistics for each analysis were obtained (Table 2). Except for the Rasch analysis, the reliability of both tests after removing misfitting items reached reasonably acceptable levels of internal consistency (i.e., α>.70) for all models.

In both tests, the CTT and IRT 2PL methods produced the highest internal estimates of consistency among items. It is worth noting that the total score for a test that assesses many different topics (e.g., the mid-term test), on which the standard error of measurement depends, is likely to have a lower correlation between the total and any single content-focused item. Nonetheless, a heterogeneous item pool in terms of content may still be necessary to sample the intended domain of the test. Hence, it may be unwise to place too many eggs in the basket of high internal estimates of reliability when evaluating a test aiming to cover multiple topic areas. However, this seems not to be a problem in this instance, if either the CTT or IRT 2PL models are used to remove poor fitting items, and less so if the IRT 2PL approach is used because it retained a greater number of items than the CTT approach.

<insert Table 2 about here>

Grade Impact

After removal of misfitting items, the grade distribution for the midterm and final examinations and the cumulative effect of the model changes on total course grade were determined for the IRT 2PL model only (Table 3). Interrater agreement using Fleiss' (1971) generalized Kappa (κ) was conducted to evaluate the chance-corrected measure of agreement between the two rating systems, each of which independently classified the subjects into one of a set of 11 (A to D-) mutually exclusive and exhaustive grade categories. King's (2004) software reported an overall κ=0.25 (95%CI=.23—.28), much below the minimum standard of κ>.40 to indicate that the observed agreement is greater than might occur by chance (Stemler, 2004). Interestingly, the kappa value per grade category was similarly low, except for the 'D-' grade.

However, when aggregated into the four main grade categories (A to D), the proportion agreement was 70%, giving κ=0.55, a somewhat more convincing indication that grade similarity was beyond chance. Grade results changed for nearly two-thirds of the 372 participants, nearly equally split between increase (*n*=123) and decrease (*n*=123). The total number of 'A' grades increased trivially from 24 to 26; 'B' grades fell from 146 to 133; the number of 'C' grades increased from 142 to 149; and the number of fail grades increased from 60 to 64. Hence, it could be argued that using the 2PL IRT approach would not make the course look any worse in terms of grade distribution because there were two more 'A's and only four more fails.

<insert Table 3 about here>

Discussion

This study showed that the instructor-constructed MCQs used in this higher education course were problematic, much more so in the midterm test than the final examination. The inclusion of poor quality items had a small but critical impact on students' overall course grades, especially in terms of pass/fail decisions. Using a statistical model approach to removing items with unacceptable characteristics made a difference to course performance in a way that benefitted a small number of students and overall made the course appear equally successful as the official raw score approach.

Given that MCQs have the possibility of guessing, it seems logically appropriate to analyse items with a statistical model capable of detecting the effect of chance performance. This is a feature only of the IRT 3PL statistical model. However, the current study has shown that the IRT 2PL model had superior fit to the overall data, indicating that this analysis can be sufficient to detect items with low or reverse discrimination, leading to appropriate calculation of person ability, and ultimately an appropriate grade score. It may be that the pseudo-chance guessing parameter cannot be accurately estimated with *N*< 1000, and so the IRT 2PL may fit better simply because of the relatively low *N* in this operational test. Further evaluation of model fit in large enrolment classes (e.g., *N*>1000) could be conducted with operational MCQ tests in many large universities.

In contrast, the IRT 1PL model did the worst job, especially in the midterm test, when perhaps item writing quality was weaker than the final examination**.** The strict assumptions of the Rasch model seem to be unrealistic for use with MCQs (and quite possibly all dichotomously scored knowledge questions) and so this analysis reiterates previous findings. It may be that when items are written better or when students make greater effort, both of which are possible explanations for the better properties of the items in the final examination, the IRT 2PL model may be sufficient (Crocker & Algina, 1986).

Given the much simpler statistical manipulations involved in calculating a CTT score relative to any of the IRT models, it may be tempting to conclude that it is sufficiently robust. However, to illustrate the additional benefit of using an IRT approach over the CTT method, consider two students (i.e., AUID153 and AUID332) who both answered correctly 15 of 34 items on the CTT revised final exam for a percentage score of 44%. This is considered an unsatisfactory grade showing a lack of knowledge and understanding of the topic. However, after adjusting student scores based on the relative difficulty of items using the IRT 2PL method, student AUID153 would get four points less moving their total grade 'C' (i.e., satisfactory), while student AUID332 would get 0.6 points more resulting in grade 'D-' (i.e., extremely poor). This suggests that insofar as the MCQ items were concerned, student AUID332 answered a few harder items correctly relative to AUID153. Thus, treating items of different difficulty as if they have the same information about the quality of performance would in these two cases might generate different conclusions. Nonetheless, since half the course grade depended on performance on essay-type questions, this small change in test score was not sufficient to change total grade.

At the same time, using a test with many fewer items may also be misleading, since the abbreviated test is likely to cover a much smaller part of the intended curriculum. Thus, the statistical analysis might lead to a more credible score, it may do so at the cost of valid inferences about competence across the full range of examination objectives. Given the power of the IRT 2PL model to adjust scores based on the relative difficulty of items, a case could be made for retaining poor quality items to ensure content coverage. However, it is our position that having fewer items would lead to more defensible decision making than retaining poor quality items that generate misleading information. Fewer items with trusted information can result in more robust decision than poorly constructed items. In terms of a purely formative assessment that does not contribute to summative grading, maximising content coverage may be useful, but when coursework and class quizzes or tests contribute to final total grades, we are of the opinion that making that overall judgement based on high-quality information is more likely to lead to public credibility.

Not using an IRT approach to score calculation potentially has a negative impact on instructors. For example, initial raw scores showed that 217 of 375 students had failed the midterm test, whereas, if IRT 2PL had been used only 144 students would have failed, a nearly 20% decline in the fail rate. Since a high proportion of failing grades can be interpreted as poor teaching quality (Brown, 2010), the raw test information in this situation may have led instructors to invest time and resources to changing teaching strategies, which may not have been needed or could have been spent in a more productive way. Thus, not conducting item analysis and removing poorly performing items could lead to misleading feedback to both students and lecturers. Nonetheless, given the superior fit of the IRT 2PL model to the data, it may be that our logical preference for the IRT 3PL is misplaced empirically and greater emphasis should be put on using the simpler statistical model is sufficient.

Alternatively, the poor characteristics of the mid-term test items may suggest a different explanation. If a test is too difficult, it is understandable many students would guess. Difficulty for the students may arise from poorly written items, but also from poor instruction. However, this was not the first time the course had been run and the course content and sequences followed those set by previous administrations of the course. The only difference to previous administrations of the course was a different teacher for the first five lectures and, thus, a different item writer for the mid-term test. Hence, it seems unlikely that the present study has identified a need to revise the course. Rather it seems more likely that there was a greater need for item analysis of the MCQ mid-term test before scores were finalised.

The overwhelming conclusion is that item statistical analysis is a necessary adjunct to judgment-based evaluation of item quality in MCQ testing in higher education. The quality of decisions can be defended when the statistical analysis eliminates misleading items. However, this requires that before any scores are released to students or record systems, some sort of psychometric analysis of item characteristics has to be conducted. Since most higher education teachers would have little training in these procedures, it seems that the development of automated analysis systems would be a useful support for academics. An automated system would indicate that certain items do not meet statistical conventions, with an opportunity for the academic to approve deletion. Once poorly fitting items are deleted, the system would recalculate scores for students. By displaying the items from easiest to hardest, the system could then ask the instructor, using the logic of bookmark standard setting (Mitzel, Lewis, Patz, & Green, 2001), to indicate where boundaries for each grade level should be established. Having done this, the system would then transform IRT logit scores into appropriate institutional values reflecting the grade boundary decisions made by the academic. For example, a score of -0.15 might be judged to be the minimum passing mark of

50%. Then actual grades could be stored in student management systems and disseminated to students. This approach takes advantage of computer technology to calculate scores while placing responsibility for grade boundaries in the hands of the content experts teaching a course (Pitoniak & Morgan, 2012).

Future Research

This last possibility identifies a clear weakness in this study. No expert based judgement of the revised test items was conducted to determine appropriate cut scores for grades. Instead, the distributions were transformed to match the raw score distribution, which had not been subjected to a standard setting process. Another challenge to conducting IRT analysis is the availability of open-source software suitable for these analyses. While SPSS has developed routines for 1PL and 2PL analysis, the base product is not free. Likewise, Mplus, which only provides 2PL analyses, is not free. The ICL (Hanson, 2002) and PARAM (Rudner, 2012) applications are free and can run IRT 3PL analyses, but are not widely used. The free ‘ltm’ package (Rizopoulos, 2006) in R overcomes these challenges. While very similar conclusions about which items to keep or reject would be reached across the various applications, there were more than trivial differences in the difficulty and discrimination parameters, for example between ICL and PARAM. Hence, analysts would be greatly aided by studies which can establish equivalences between open source applications and gold-standard applications. Perhaps, with greater acceptance and use, appropriate packages (e.g., ‘ltm’ or ‘mirt’) in the open source R software will be able to resolve these equivalence and access problems.

Another limitation in this study is the relatively small sample size (i.e., 380 students) and the effect it may have on estimating the pseudo-chance parameter in the IRT 3PL model. However, small sample size is a normal case in higher education and it may be relatively uncommon to have courses with at least 1000 students. Nonetheless, future studies to further establish the robustness of IRT applications for realistically small sample sizes are needed.

The generalizability of this study is limited because only one course and only two MCQ tests in one year have been analysed. Nonetheless, the current study is consistent with other studies that have evaluated the quality of multiple-choice items in higher education assessments. Thus, further studies into the quality of MCQ testing, especially evaluating training programs designed to improve instructor item writing skill, are needed. Evidence from publishers is needed about the qualities of items in text-book related item banks which can be used in formal assessments.

Conclusion

This research is necessary within each and every institution that uses MCQs, because while the threat cumulatively may not be large, it seems highly likely that specific exams or tests will not meet normal requirements. The credibility of assessment is necessary, especially if there is a tendency to be litigious about grading or testing (Brookhart, 2009). Any lack of quality assurance process at the course or department level poses a significant reputational risk to the institution.

This study has shown that use of IRT item analysis has a potential beneficial impact on overall course grades and number of students passing. It also suggests that more informative feedback to students and instructors might be generated by giving grades based derived from item difficulty. This study provides a warning for the different stakeholders concerned with the quality of higher education assessment practices and suggests that more commitment and effort is needed in quality assurance in order to meet professional obligations.

References

Adedoyin, O. & Adedoyin, J. (2013).Assessing the comparability between classical test theory (CTT) and item response theory (IRT) models in estimating test item parameters. *Herald Journal of Education and General Studies, 2*(3), 107-114.

Adedoyin, O., & Mokobi, T. (2013). Using IRT psychometric analysis in examining the quality of junior certificate mathematics multiple choice examination test items. *International Journal of Asian Social Science, 3*(4), 992-1011.

Akour, M., & AL-Omari, H. (2013). Empirical investigation of the stability of IRT item-parameters estimation. *International Online Journal of Educational Sciences, 5*(2), 291-301.

Athanasou, J. A., & Lamprianou, I. (2004). Reading in one's ethnic language: A study of Greek-Australian high school students. *Australian Journal of Educational & Developmental Psychology, 4*, 86-96.

Bailey, P. H., Mossey, S., Moroso, S., Cloutier, J. D., & Love, A. (2012). Implications of multiple-choice testing in nursing education. *Nurse Education Today, 32*(6), e40-e44. doi:10.1016/j.nedt.2011.09.011

Bergan, J. (2010). *Assessing the relative fit of alternative item response theory*. Tucson, Arizona: Assessment Technology Incorporated.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch Analysis in the Human Sciences*. Dordrecht, NL: Springer Science.

Borsboom, D. (2005). Measuring the mind conceptual issues in contemporary psychometrics. Cambridge, UK: Cambridge University Press.

Brady, A. (2005). Assessment of learning with multiple-choice questions. *Nurse Education in Practice, 5*(4), 238-242. doi:10.1016/j.nepr.2004.12.005

Brookhart, S. M. (2009, updated 20 July, 2010). Legal context of grading. Retrieved 25 January, 2017, from https://www.education.com/reference/article/legal-context-grading/

Brown, G. T. L. (2010). The validity of examination essays in higher education: Issues and responses. *Higher Education Quarterly, 64*(3), 276-291.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection. Sociological Methods & Research, 33(2), 261-304. 10.1177/0049124104268644

Burton, R. F. (2001). Quantifying the effects of chance in multiple choice and true/false tests: Question selection and guessing of answers. *Assessment & Evaluation in Higher Education, 26*(1), 41-50. doi:10.1080/02602930020022273

Burton, R. F. (2005). Multiple-choice and true/false tests: Myths and misapprehensions. *Assessment & Evaluation in Higher Education, 30*(1), 65-72. doi:10.1080/0260293042003243904

Cizek, G. J. (Ed.). (2001). *Setting Performance Standards: Concepts, Methods, and Perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates.

Clifton, S. L., & Schriner, C. L. (2010). Assessing the quality of multiple-choice test items. *Nurse Educator, 35*(1), 12-16.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston.

de Ayala, R. J. d. (2009). *The theory and practice of item response theory*. New York: The Guilford Press.

DiBattista, D., & Kurzawa, L. (2011).Examination of the quality of multiple-choice items on classroom tests. *The Canadian Journal for the Scholarship of Teaching and Learning*, *2*(2), [Article 4]. http://dx.doi.org/10.5206/cjsotl-rcacea.2011.2.4

Divgi, D. R. (1986). Does the Rasch model really work for multiple choice items? Not if you look closely. *Journal of Educational Measurement, 23*(4), 283-298.

Downing, S. M. (2006). Twelve steps for effective test development. In S. M. H. Downing, M. Thomas (Eds.), *Handbook of test development* (pp. 3-25). Mahwah, NJ: Lawrence Erlbaum Associates.

Downing, S. M., & Yudkowsky, R. (2009).*Assessment in health professions education*. New York, NY: Routledge.

Downing, S.M. (2005). The effects of violating standard item-writing principles on tests and students: The consequences of using flawed test items on achievement examinations in medical education. *Advances in Health Sciences Education, 10*(2), 133–43.

Drasgow, F., & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement, 7*(2), 189–199.

Ebel, R. L.,& Frisbie, D. A. (1991). *Essentials of educational measurement* (5thed.). Englewood Cliffs, NJ: PrenticeHall.

Ellsworth, R. A., Dunnell, P., & Duell, O. K. (1990). Multiple-choice test items: What are textbook authors telling teachers? *The Journal of Educational Research*, *83*(5), 289-293.

Embretson, S., & Reise, S. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum Associates.

Fan, X., & Sivo, S. A. (2007). Sensitivity of fit indices to model misspecification and model types. *Multivariate Behavioral Research, 42*(3), 509–529. doi: 10.1080/00273170701382864

Finney, S. J., & DiStefano, C. (2006). Non-normal and categorical data in structural equation modeling. In G. R. Hancock & R. D. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 269-314). Greenwich, CT: Information Age Publishing.

Fellenz, M. R. (2004). Using assessment to support higher level learning: The multiple choice item development assignment. *Assessment & Evaluation in Higher Education, 29*(6), 703-719. doi:10.1080/0260293042000227245

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*(5), 378-382. doi:10.1037/h0031619

Giblin, P. J. (1972). What is an asymptote? *The Mathematical Gazette, 56*(398), 274-284.

Grainger, P., Purnell, K., & Zipf, R. (2008). Judging quality through substantive conversations between markers. *Assessment & Evaluation in Higher Education*, *33*(2), 133-142.

Haladyna, T. (2004). *Developing and validating multiple-choice test items*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Haladyna, T. M., & Downing, S. M. (1993). How Many Options is Enough for a Multiple-Choice Test Item? *Educational and Psychological Measurement, 53*(4), 999-1010.

Haladyna, T. R., M. & Rodriguez, M. C. (2013). *Developing and validating test items*. New York: Routledge Taylor & Francis Group.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, *12*(3), 38-47.

Hambleton, R., Swaminathan, H., & Rogers, H. (1993).*Fundamentals of Item Response Theory*. California: Sage Publications.

Hansen, J. D., & Dexter, L. (1997). Quality multiple-choice test questions: Item-writing guidelines and an analysis of auditing test banks. *Journal of Education for Business, 73*(2), 94-97. doi:10.1080/08832329709601623

Hanson, B. A. (2002). ICL: IRT command language (Version 0.020301) [Computer software]. Retrieved from http://www.b-a-h.com/software/irt/icl/

Hingorjo, M. R., & Jaleel, F. (2012). Analysis of One-Best MCQs: the Difficulty Index, Discrimination Index and Distractor Efficiency. *Journal of the Pakistan Medical Association, 62*(2), 142-147.

Houts, C. R., Edwards, M. C., Wirth, R. J., & Deal, L. S. (2016). A review of empirical research related to the use of small quantitative samples in clinical outcome scale development. *Quality of Life Research, 25*, 2685-2691. doi:10.1007/s11136-016-1364-9

Joughin, G. (2009). Assessment, learning and judgment in higher education: A critical review. In G. Joughin (Ed.), *Assessment, learning and judgment in higher education* (pp. 1-15).Dordrecht, Netherlands: Springer.

Jozefowicz, R. F., Koeppen, B. M., Case, S., Galbraith, R., Swanson, D., & Glew, R. H. (2002). The quality of in-house medical school examinations. *Academic Medicine, 77*(2), 156-161.

King, J. E. (2004, 5-7 February). *Software solutions for obtaining a kappa-type statistic for use with multiple raters*. Paper presented at the annual meeting of the Southwest Educational Research Association, Dallas, TX.

Leeson, H., & Fletcher, R. (2003).*An investigation of fit: Comparison of 1-, 2-, 3-parameter IRT models to project asTTle data*. Paper presented at the Joint NZARE/AARE Conference, Auckland, New Zealand.

Malau-Aduli, B. S., & Zimitat, C. (2012). Peer review improves the quality of MCQ examinations. *Assessment & Evaluation in Higher Education*, *37*(8), 919-931.

Masters, J. C., Hulsmeyer, B. S., Pike, M. E., Leichty, K., Miller, M. T., & Verst, A. L. (2001). Assessment of multiple-choice questions in selected test banks accompanying text books used in nursing education. *Journal of Nursing Education, 40*(1), 25-32.

Mitzel, H.C., Lewis, D.M., Patz, R.J., & Green, D.R. (2001). The bookmark procedure: Psychological perspectives. In G.J. Cizek (Ed.), S*etting performance standards: Concepts, Methods, and Perspectives* (pp. 249-281). Mahwah, NJ: Erlbaum.

Muthén, L. K., & Muthén, B. O. (1998-2015). *Mplus User's Guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Nedeau-Cayo, R., Laughlin, D., Rus, L., & Hall, J. (2013).Assessment of item-writing flaws in multiple-choice questions. *Journal for Nurses in Professional Development*, *29*(2), 52-57.

Osterlind, S., & Wang, Z. (2012).Item response theory in measurement, assessment, and evaluation for higher education. In Secolsky, C., & Denison, D. (Eds.)*, Handbook on Measurement, Assessment, and Evaluation in Higher Education* (pp. 150-160). New York, NY: Routledge

Pitoniak, M. J., & Morgan, D. L. (2012). Setting and validating cut scores for tests. In C. Secolsky, & D. B. Denison (Eds.), *Handbook on Measurement, Assessment, and Evaluation in Higher Education* (pp. 343-366). New York: Routledge.

Popham, W. J. (2011). *Classroom assessment: What teachers need to know*. Boston, MA: Pearson.

Raykov, T., & Marcoulides, G. (2006). *A first course in structural equation modeling*. Mahwah, NJ: Lawrence Erlbaum Association.

Reynolds, C. R., Livingston, R. B., & Wilson, V. (2009). *Measurement and Assessment in Education*. Upper Saddle River, NJ: Pearson Merrill.

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response analysis. Journal of Statistical Software, 17(5), 1-25. doi:10.18637/jss.v017.i05

Rudner, L. M. (2012). PARAM calibration software logistic IRT model (Version 0.93) [Computer software]. Retrieved from http://echo.edres.org:8080/irt/param/

Schaughency, E., Smith, J. K., van der Meer, J., & Berg, D. (2012). Classical test theory and higher education: Five questions. In C. Secolsky & D. B. Denison (Eds.), *Handbook on Measurement, Assessment, and Evaluation in Higher Education* (pp. 117-131). London: Routledge.

Sireci, S. G. (1992. *The Utility of IRT in Small-Sample Testing Applications*. Paper presented at the Annual Meeting of the American Psychological Association, Washington, DC

Stagnaro-Green, A. S., & Downing, S. M. (2006). Use of flawed multiple-choice items by the New England Journal of Medicine for continuing medical education. *Medical Teacher, 28*(6), 566-568.

Tarrant, M., & Ware, J. (2008).Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. *Medical Education, 42*(2), 198-206. doi:10.1111/j.1365-2923.2007.02957.x

Tarrant, M., Knierim, A., Hayes, S. K., & Ware, J. (2006). The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse Education in Practice, 6*(6), 354-363.

Tarrant, M., Ware, J., & Mohammed, A. M. (2009). An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis. *BMC Medical Education, 9*(40) doi:10.1186/1472-6920-9-40

Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 73-136). Mahwah, NJ: Lawrence Erlbaum Associates.

van de Vijver, F. J. (1986). The robustness of Rasch estimates. *Applied Psychological Measurement*, *10*(1), 45-57.

Walsh, C. M., & Seldomridge, L. A. (2006). Critical thinking: Back to square two. *The Journal of Nursing Education, 45*(6), 212-219.

Walvoord, B. E., & Anderson, V. J. (2011). *Effective grading: A Tool for Learning and Assessment in College*. San Francisco: Wiley.

Wise, S. L., & Smith, L. F. (2016). The validity of assessment when students don't give good effort. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of Human and Social Conditions in Assessment* (pp. 204-220). New York: Routledge.

Yorke, M. (2009). Faulty signals? Inadequacies of grading systems and a possible response. In G. Joughin, (Ed.) *Assessment, Learning and Judgement in Higher Education* (pp. 1-20). London, UK: Springer.

Yu, C.-Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes.* (Unpublished doctoral dissertation), University of California, Los Angeles, Los Angeles, CA.

Table 1. Item psychometric characteristics by statistical model and test

Item	CTT			IRT 1PL RASCH		IRT 2PL		IRT 3PL		
	p value (% correct)	Point biserial correlation (r_{pb})	DE	Difficulty (b)	χ^2 *p*	Difficulty (b)	Discrimination (a)	Discrimination (a)	Difficulty (b)	Chance (c)
Midterm test										
M1	78.70	-.06	0.67	-0.90	0.405	-3.16	0.36	0.21	-4.00	0.00
M2	39.20	.47	1.00	-0.96	0.035	0.12	-6.62	0.76	-1.30	0.01
M3	37.90	.25	1.00	-0.07	0.030	-0.10	-0.95	7.25	1.86	0.50
M4	55.20	.04	1.00	-1.13	0.220	0.28	1.16	0.48	-2.28	0.00
M5	57.10	.01	1.00	-1.33	0.483	0.20	1.19	0.43	-2.66	0.10
M6	18.70	-.20	1.00	2.33	0.022	2.69	0.75	-0.26	-8.47	0.00
M7	61.30	.11	0.67	-1.66	0.003	0.03	1.22	1.19	-1.62	0.00
M8	4.30	.49	1.00	-1.05	0.027	0.15	-6.24	0.97	-1.19	0.00
M9	.80	-.01	0.67	-1.75	0.203	14.10	0.36	1.40	0.10	0.69
M10	40.00	.54	1.00	-0.93	0.225	0.14	-6.71	0.75	-1.28	0.00
M11	20.00	-.06	1.00	1.20	0.000	4.10	0.40	-0.36	-3.12	0.00
M12	14.40	-.13	0.67	1.68	<0.0001	6.74	0.29	-0.64	-2.62	0.00
M13	81.90	.06	1.00	-1.51	0.002	-2.74	0.48	1.14	-1.52	0.00

Item	CTT		IRT 1PL RASCH			IRT 2PL		IRT 3PL		
	p value (% correct)	Point biserial correlation (r_{pb})	DE	Difficulty (b)	χ^2 *p*	Difficulty (b)	Discrimination (a)	Discrimination (a)	Difficulty (b)	Chance (c)
M14	66.40	.00	0.67	-2.53	0.037	-0.08	1.73	9.88	0.44	0.87
M15	47.20	-.02	1.00	-0.69	0.013	0.65	0.89	1.66	0.46	0.45
M16	44.30	.54	1.00	-1.75	0.007	0.09	-18.15	1.16	-1.73	0.00
M17	39.70	-.14	1.00	0.63	0.170	1.48	0.46	0.35	2.02	0.04
M18	36.80	.49	1.00	-1.02	0.219	0.10	-4.76	2.62	0.95	0.64
M19	4.80	.49	1.00	-1.34	0.051	0.13	-8.73	9.81	0.47	0.66
M20	85.90	.08	0.67	-1.82	0.004	-2.92	0.55	1.08	-1.89	0.00
M21	40.00	-.41	1.00	-0.17	0.014	0.85	1.51	26.98	1.13	0.46
M22	62.40	.45	1.00	-1.19	0.212	0.89	-1.77	1.38	-0.22	0.44
M23	1.30	-.02	0.33	-0.87	0.203	-26.93	-0.16	11.55	0.98	0.64
M24	49.90	.32	0.67	-0.54	0.357	0.48	-1.29	0.56	-0.53	0.13
M25	59.70	-.21	1.00	-0.01	0.324	0.08	1.14	1.73	1.22	0.39
M26	34.90	.00	1.00	0.17	0.503	1.41	0.78	37.50	1.40	0.43
M27	35.70	.46	1.00	-0.70	0.042	0.08	-4.97	0.82	-0.92	0.00
M28	45.60	.08	1.00	-0.50	0.107	0.77	0.76	0.60	-0.86	0.00

	CTT			IRT 1PL RASCH		IRT 2PL		IRT 3PL		
Item	*p* value (% correct)	Point biserial correlation (r_{pb})	DE	Difficulty (b)	χ^2 *p*	Difficulty (b)	Discrimination (a)	Discrimination (a)	Difficulty (b)	Chance (c)
M29	66.10	-.14	1.00	-0.52	0.034	-0.97	0.46	0.04	7.68	0.34
M30	65.90	.47	1.00	-2.06	0.035	0.90	-2.61	1.75	-0.53	0.63
M31	44.50	.47	0.67	-1.51	0.084	0.19	-9.09	0.98	-1.36	0.19
M32	44.80	.28	1.00	-0.07	0.001	0.26	-1.03	0.87	0.13	0.09
M33	53.90	-.27	1.00	0.27	0.004	0.32	1.03	-0.17	-1.37	0.00
M34	63.50	.05	0.67	-2.03	0.001	0.01	1.64	1.67	-1.61	0.00
M35	39.50	.45	1.00	-0.74	0.085	0.14	-5.81	16.16	1.46	0.64
M36	67.20	.49	0.67	-1.84	0.004	0.96	-2.49	1.50	-1.55	0.00
M37	44.30	.56	0.67	-1.98	0.001	0.18	-8.89	1.24	-1.87	0.00
M38	45.30	.54	0.67	-1.91	0.015	0.06	-30.03	1.21	-1.83	0.00
M39	35.20	.07	1.00	0.16	0.272	1.56	0.64	0.18	0.75	0.00
M40	68.30	-.03	0.33	-2.49	0.650	-0.05	2.61	0.88	-2.97	0.00
M41	64.80	.49	0.67	-1.86	0.069	0.89	-2.35	1.06	-1.96	0.00
M42	81.30	.00	0.67	-1.54	0.421	-10.04	0.14	0.33	0.00	0.61
M43	4.00	.49	1.00	-1.27	0.271	0.14	-6.21	0.54	-2.34	0.00

Item	CTT		IRT 1PL RASCH			IRT 2PL		IRT 3PL		
	p value (% correct)	Point biserial correlation (r_{pb})	DE	Difficulty (b)	χ^2 *p*	Difficulty (b)	Discrimination (a)	Discrimination (a)	Difficulty (b)	Chance (c)
M44	59.20	.47	1.00	-1.00	0.004	0.77	-1.82	1.06	-1.07	0.00
M45	21.90	.07	1.00	2.10	0.022	-0.23	-4.16	-0.31	-6.44	0.00
M46	47.50	-.04	1.00	-0.80	0.278	0.62	1.05	0.50	-1.56	0.00
M47	56.80	.45	1.00	-0.96	0.002	0.70	-1.72	1.22	-0.93	0.00
M48	2.80	.11	0.67	1.07	0.014	3.06	0.56	0.91	1.62	0.07
M49	53.90	.47	1.00	-0.90	0.319	0.60	-1.72	0.67	-1.37	0.00
M50	57.90	-.16	1.00	-0.09	0.287	0.00	0.66	0.60	-0.17	0.00
Final Exam										
E01	27.10	.14	1.00	1.13	0.57	2.57	0.40	0.64	2.59	0.11
E02	84.90	.36	1.00	-1.95	0.02	-1.61	1.45	2.39	-0.68	0.46
E03	46.90	.15	1.00	0.14	0.18	0.38	0.33	0.32	0.40	0.00
E04	59.90	.22	1.00	-0.46	0.97	-0.93	0.46	0.58	0.11	0.22
E05	83.60	.32	1.00	-1.84	0.21	-1.85	1.06	1.11	-1.49	0.20
E06	4.80	.20	1.00	0.42	0.02	0.80	0.48	6.64	1.26	0.34
E07	58.60	.45	1.00	-0.40	0.00	-0.34	1.54	1.92	-0.09	0.11

Item	CTT			IRT 1PL RASCH		IRT 2PL		IRT 3PL		
	p value (% correct)	Point biserial correlation (r_{pb})	DE	Difficulty (b)	χ^2 *p*	Difficulty (b)	Discrimination (a)	Discrimination (a)	Difficulty (b)	Chance (c)
E08	24.10	.09	0.67	1.30	0.27	4.78	0.24	0.49	4.02	0.12
E09	53.10	.39	1.00	-0.14	0.01	-0.15	1.13	2.02	0.32	0.21
E10	6.10	.00	1.00	3.00	0.13	82.28	0.03	2.55	3.41	0.06
E11	7.80	.26	1.00	-1.01	0.97	-1.69	0.56	0.57	-1.67	0.00
E12	63.70	.15	1.00	-0.64	0.58	-1.62	0.36	0.37	-1.54	0.00
E13	94.70	.19	0.33	-3.18	0.35	-3.49	0.93	0.88	-3.69	0.00
E14	79.60	.27	1.00	-1.54	0.60	-1.85	0.84	2.13	0.07	0.60
E15	59.40	.19	1.00	-0.44	0.02	-0.77	0.53	19.80	0.77	0.51
E16	95.20	.16	0.67	-3.30	0.15	-4.15	0.79	0.71	-4.53	0.01
E17	47.70	.13	1.00	0.10	0.52	0.26	0.36	0.64	1.51	0.26
E18	56.00	.28	1.00	-0.28	0.28	-0.40	0.67	0.67	-0.39	0.00
E19	69.20	.11	1.00	-0.93	0.01	-3.59	0.23	0.27	-3.07	0.00
E20	97.90	.09	0.33	-4.16	0.70	-6.19	0.65	0.62	-6.43	0.05
E21	82.00	.53	0.67	-1.71	<0.0001	-1.14	2.42	2.26	-1.13	0.00
E22	55.40	.37	1.00	-0.25	0.01	-0.26	1.09	1.28	0.06	0.13

	CTT			IRT 1PL RASCH		IRT 2PL		IRT 3PL		
Item	*p* value (% correct)	Point biserial correlation (r_{pb})	DE	Difficulty (b)	χ^2 *p*	Difficulty (b)	Discrimination (a)	Discrimination (a)	Difficulty (b)	Chance (c)
E23	87.80	.42	0.67	-2.22	0.01	-1.67	1.71	1.67	-1.58	0.10
E24	66.30	.34	1.00	-0.77	0.33	-0.94	0.84	1.15	-0.09	0.29
E25	51.20	.34	1.00	-0.06	0.24	-0.07	0.94	1.17	0.33	0.15
E26	63.10	.22	1.00	-0.62	0.38	-1.07	0.54	1.37	0.83	0.47
E27	63.70	.25	1.00	-0.64	0.40	-0.95	0.65	1.53	0.65	0.46
E28	85.70	.41	0.67	-2.02	0.00	-1.58	1.58	1.93	-1.06	0.32
E29	72.40	.40	1.00	-1.10	0.07	-1.03	1.21	1.15	-1.04	0.00
E30	82.80	.34	0.67	-1.77	0.51	-1.72	1.13	1.28	-1.06	0.33
E31	21.50	.24	1.00	1.46	0.07	1.88	0.77	1.61	1.62	0.09
E32	78.20	.36	1.00	-1.45	0.15	-1.39	1.15	1.31	-0.94	0.22
E33	87.50	.47	0.67	-2.19	<0.0001	-1.58	1.86	1.74	-1.63	0.00
E34	74.50	.13	1.00	-1.22	0.15	-2.89	0.39	1.51	0.87	0.64
E35	77.50	.42	1.00	-1.40	0.01	-1.22	1.36	1.32	-1.22	0.00
E36	13.80	-.10	1.00	1.39	0.00	-70.25	-0.02	-0.07	-19.63	0.02
E37	73.20	.36	1.00	-1.15	0.02	-1.22	0.99	0.94	-1.25	0.00

Item	CTT		IRT 1PL RASCH			IRT 2PL		IRT 3PL		
	p value (% correct)	Point biserial correlation (r_{pb})	DE	Difficulty (b)	χ^2 *p*	Difficulty (b)	Discrimination (a)	Discrimination (a)	Difficulty (b)	Chance (c)
E38	82.00	.27	0.67	-1.71	0.62	-2.18	0.78	0.76	-2.21	0.00
E39	52.30	.32	1.00	-0.10	0.40	-0.15	0.72	0.87	0.17	0.10
E40	5.40	.27	1.00	-0.02	0.19	-0.03	0.63	1.24	0.74	0.26
E41	68.40	.22	1.00	-0.88	0.15	-1.39	0.60	1.94	0.64	0.54
E42	93.90	.12	0.67	-3.03	0.92	-5.65	0.50	0.50	-5.64	0.00
E43	72.10	.20	1.00	-1.09	0.13	-2.09	0.48	0.52	-1.95	0.00
E44	17.20	.04	0.33	4.01	0.00	-4.01	-1.05	-0.99	-4.23	0.00
E45	71.90	.37	1.00	-1.07	0.12	-1.10	1.05	1.04	-1.08	0.00
E46	76.70	.29	1.00	-1.35	0.17	-1.58	0.87	2.53	0.16	0.57
E47	85.70	.45	0.67	-2.02	0.00	-1.46	1.89	1.83	-1.37	0.09
E48	89.10	.39	0.67	-2.36	0.01	-1.69	1.87	2.10	-1.37	0.25
E49	81.40	.47	0.67	-1.68	<0.0001	-1.25	1.80	2.39	-0.77	0.28
E50	19.40	-.02	1.00	1.61	0.00	-68.59	-0.02	-0.02	-74.32	0.06

Note. DE=discriminator efficiency; Items marked in red were rejected as not meeting model conventions.

Table 2. Midterm Test and Final Exam Statistics by Statistical Model

	Midterm test					Final Exam				
Statistic	Raw	CTT	Rasch	2PL	3PL	Raw	CTT	Rasch	2PL	3PL
k	50	22	24	26	30	50	34	31	47	32
M	23.78	10.53	11.36	13.02	15.69	32.10	21.76	19.95	31.66	20.54
SD	5.96	7.40	3.10	5.07	4.94	6.64	5.99	4.10	6.64	4.50
SEM	3.21	1.81	2.21	2.09	2.52	2.82	2.40	2.28	2.74	1.86
Cronbach α	0.71	0.94	0.49	0.83	0.74	0.82	0.84	0.69	0.83	0.83

Note. *k*=number of items

Table 3. Revised and Original Grade Distributions

	Revised Grade											Original Total
Original Grade	A	A-	B+	B	B-	C+	C	C-	D+	D	D-	
A	**1**	*1*	*2*									4
A-	4	**8**	*2*	*3*	*3*							20
B+	2	4	**11**	*11*	*5*	*4*						37
B		4	10	**14**	*9*	*7*	*4*					48
B-		2	4	13	**21**	*12*	*9*					61
C+				7	11	**16**	*12*	*7*	*4*			57
C					5	15	**23**	*4*	*6*			53
C-					2	3	10	**10**	*2*	*5*		32
D+						1	2	6	**6**	*3*	*4*	22
D							2	1	6	**6**	*3*	18
D-							1		7	1	**11**	20
Revised Total	7	19	29	48	56	58	63	28	31	15	18	372
Fleiss' Generalised Kappa	.17	.38	.27	.19	.24	.15	.29	.27	.17	.33	.56	.25

Note. Bold indicates number of cases who would receive the same grade. Values in italic in the top half of diagonal indicate cases who receive lower grades after revision. Values in bottom half indicate cases who would receive higher grades after revision.