An Integrated Methodology for Assessing Item Discrimination in Mathematics Assessments

Russell Jeter[a,b,†*], Darryl Chamberlain, Jr.[c,†], and Kelvin Rozier[a]
[a] Department of Mathematics and Statistics, Georgia State University, P.O. Box 4110, Atlanta, Georgia, 30302-410, USA
[b] Neuroscience Institute, Georgia State University, P.O. Box 4110, Atlanta, Georgia, 30302-410, USA
[c] Department of Mathematics, Science and Technology, Embry-Riddle Aeronautical University, 1 Aerospace Boulevard, Daytona Beach, FL 32114-3900, USA
[†]These authors contributed equally to this work
[*]Corresponding author: rjeter@gsu.edu

Abstract

In order to assess an assessment, one must consider whether the assessment produces valid and reliable results according to its intended goals. Given the various ways that validity and reliability can be assessed based on the goals of some assignment, there is no singular methodology for analyzing an educational assessment. This study attempts to address this research gap by presenting a unifying methodology that incorporates both CTT and IRT, along with other measures of validity and reliability, that can provide a robust analysis of an educational assessment while considering the goals of that assessment. Utilizing item measures from Classic Test Theory, Item Response Theory, and Distractor Analysis, we present a summary of each metric that will inform labeling items as sufficiently discriminating for the purposes of an educational assessment. We then present an analysis of 4 exams to illustrate how the metrics can be used in concert to identify validity and reliability at the item and assessment level. We conclude with a brief discussion of how this methodology can be applied to different types of assessments.

Keywords: assessment; classical test theory; item response theory; distractor analysis

Introduction

Assessment is an integral part of every course. While all assessments attempt to measure student knowledge, the inherent goals of each assignment drives the design. When one hears the word "assessment" they often consider high-stakes exams that attempt to evaluate student knowledge against some predefined metrics. Final exams, placement exams, and general standardized tests like ACT and SAT all fit this category, and fall under *evaluative assessments*. In contrast, *formative assessments* attempt to measure student knowledge to further enhance the student's learning. This can be further split by the goal of the formative assessment: assessment as learning (AaL) and assessment of learning (AoL). In AaL, the process of completing the assessment acts as the learning opportunity. Homework and group work normally serve the goals of AaL. In AfL, the instructor uses evidence collected during assessments to guide further instruction. In AoL, the assessment attempts to identify whether learning has occurred and if interventions need to be

implemented. Quick surveys (such as through clicker questions) or quizzes normally serve the goals of AoL (Dann, 2014).

In order to assess an assessment, one must consider whether the assessment produces valid and reliable results according to its intended goals. Within assessment, *validity* refers to the extent a test measures what it purports to measure and *reliability* refers to the extent results can be replicated (Wu et al., 2016). For example, the goal of a placement exam is to classify students into the highest course they could take with success. Measuring the validity of a placement test would include analyzing the course grades for students placed and determining the impact of the placement test. Measuring the reliability of a placement test would include checking that students are consistently classified into some course. In contrast, the goal of a standardized test like the SAT is to measure some academic ability and report it in an ordered form to distinguish between high-performing and low-performing students. Measuring the validity of a standardized test may include checking the validity of individual questions (to ensure answering the question correctly relates to the student having some knowledge) while measuring reliability may include comparing correct response rates across versions within the same year and between different years.

Given the various ways that validity and reliability can be assessed based on the goals of some assignment, there is no singular methodology for analyzing an educational assessment. In general, there are two broad frameworks for studying assessments: Classical Test Theory (CTT) and Item Response Theory (IRT). CTT presents an analysis of the results of an assessment focusing on the overall scores on the assessment, whereas IRT presents an analysis that relates a latent, or unobservable, state (such as student ability) to an observed outcome (such as an answer choice on an assessment item). The latent variable is estimated by a series of observed outcomes as well as the relation of these outcomes to others' outcomes (estimated item difficulties).

These frameworks need not be disparate, and can in fact inform each other. As Wu et al. state in *Educational Measurement for Applied Researchers*:

> *…while there are theoretical differences between IRT and CTT…, in practice, both IRT and CTT help us with building a good measuring instrument. Consequently, CTT and IRT should be used hand-in-hand in a complementary way, and one should not discard one approach for another (p.26).*

This study attempts to address this research gap by presenting a unifying methodology that incorporates both CTT and IRT, along with other measures of validity and reliability, that can provide a robust analysis of an educational assessment while considering the goals of that assessment.

The outline of this paper is as follows: in the Theoretical Framework section, we present measures associated with CTT, IRT, and other measures for validity and reliability that will make up our unifying methodology for assessing educational assessments; in the Methods section, we frame the context with which we test and attempt to validate the theoretical framework; in the Results section, we present the findings of each measure outlined in the theoretical framework; in the Discussion sections, we relate the findings of the results for each measurement to present a

unified analysis of the discriminatory power of items in the exams studied; finally, in the Conclusions we present some ideas for utilizing this methodology and directions for future study.

Theoretical Framework

Assessment theory commonly starts with the assumption that an assessment will have options (multiple-choice) or be measured as correct/incorrect (free response). In educational assessment, multiple-choice options are constructed using plausible but incorrect options known as *distractors*. While some IRT models can accommodate distractors (such as providing partial credit for the selection of certain distractors seen as "close" to the correct answer), distractor analysis is commonly completed in an ad hoc manner. However, evaluating distractors is critical when the goal of an assessment is to enhance student learning and not to simply discriminate between high-performing and low-performing students. Therefore, our unifying methodology incorporates CTT, IRT, and distractor analysis in a coherent fashion to triangulate assessment validity and reliability.

Classical Test Theory

Classical Test Theory is the analysis of assessments based primarily on students' overall (observed) assessment scores. In general, the core assumption underlying CTT is that a student's observed score is the sum of their "true score," which measures their actual understanding of an underlying concept, and an "error score," which can result from numerous sources. The true score is defined as the expectation of the observed score, meaning that "on average" the error will cancel out of the equation. There are a few key statistics that can be measured when studying assessments using this framework: observed score and the statistical measures therein, reliability, item difficulty, and point-biserial correlation (a measure of item discrimination). A brief description for each of these key statistics follows.

Observed score distribution

One common assumption in classic psychometric is the assumption of normality in assessment results. However, educational assessments do not frequently present normal results (Ho & Yu, 2015). Thus, when presenting results on educational assessments, it is important to highlight statistics that measure the normality of data such as skew and kurtosis. Skew, an estimate of the third standardized moment of the distribution, describes where the "hump" of the data is in relation to the center (either mean or median) while kurtosis[1], an estimate of the fourth standardized moment of the distribution, describes the shape of the "tails" of the data. A normal distribution has a skew of 0 and kurtosis of 0. Early authors in non-normal assessment analysis such as Lord (1955) report that "easy" tests commonly present negative skew and that symmetric test distributions present negative kurtosis.

[1] It is common to present kurtosis as (kurtosis - 3), which is referred to as excess kurtosis as kurtosis of a normal distribution is 3.

Item difficulty

For dichotomous items (correct/incorrect), item difficulty refers to the proportion of students that answered an item correctly. For polytomous items, responses may be weighted for partial credit. For the purposes of this paper, we only consider dichotomous item difficulty. Item difficulty represents the mean score on an item. Its role in discrimination is more of an indicator that an item will not be able to be used to discriminate between high and low performing students than as an indicator of performance. An item with too high of a difficulty was answered correctly by nearly all of the students, and can thus not be used to discriminate against high performing students. The same applies for low item difficulty and low performing students. In light of this, Lord (1955) suggests ideal item difficulties as slightly higher than halfway between a random guess and 100%, with 74% and 70% for four and five option multiple choice assessment items respectively (p.189).

Point-biserial correlation

Point-biserial correlation (PBC) is an item-level statistic that measures the correlation between students' performance on a question and their performance on the exam as a whole. In other words, PBC describes how well the item discriminated between high-performing and low-performing students and is entirely distinct from item difficulty. PBC is equivalent to the Pearson correlation coefficient relating these two quantities, so the traditional interpretation of correlation coefficients can be applied: values range from -1 to 1; positive values indicate a positive correlation between the two quantities (that is, students that perform well on the exam as a whole perform well on the item), negative values indicate a negative correlation between the two quantities (students that perform well on the exam as a whole perform *poorly* on the item or students who perform poorly on the exam as a whole perform *well* on the item), and values close to zero indicate that a given question is not predictive of performance on the exam at all (and vice versa).

PBC can be computed using the following formula:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\sqrt{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}} \quad (1)$$

Where n is the number of students that completed the question being assessed, x_i is the i-th student's score on the question being assessed ($x_i = 0$ indicates a wrong answer, $x_i = 1$ indicates a correct answer), $\bar{x}$ is the mean score on the question being assessed. y_i is the i-th student's score on the other questions on the exam, $\bar{y}$ is the mean score on the other questions on the exam.

Items with negative PBC or a PBC of zero are poor assessment items as they indicate that the item is answered incorrectly by students that perform well on the exam (suggests an item has an

error) or that the item bears no relationship to student performance and can be removed without consequence. Categorization of PBCs for educational assessment purposes are not uniform as assessments primarily serve to assess the curriculum rather than as collections of items with the most predictive power (Wu et al., 2016). To motivate our categorizations of PBCs, we consider two thresholds based on the number of questions in the assessment (K) and number of students who answered questions in the assessment (n).

Our first threshold is based on the assumption of perfect determination. Given an exam with K questions, the coefficient of determination for a single question is $r^2 = \frac{1}{K}$ and thus the correlation of any one question to the overall score would be $r = \frac{1}{\sqrt{K}}$. Questions with correlations at or above $\frac{1}{\sqrt{K}}$ suggest a reasonable explanation of the overall score based on the assessment size. For example, in an exam with 10 questions, a reasonable correlation would be 0.32 versus 0.22 in an exam with 20 questions.

Our second threshold is based on Fisher's z-transformation, which converts a correlation coefficient r into a normally distributed variable z. The standard error for Fisher's z is $SE_z = \frac{1}{\sqrt{n-3}}$. By converting the correlations to a normally distributed variable, one can calculate a confidence interval with the formula $z' \pm z_{crit} * SE_z$. For significance 0.046, $z_{crit} = 2$. Thus when $z' = SE_z$, we have a confidence interval centered at 0. This suggests a reasonable poor item threshold to be $\tanh\left(\frac{1}{\sqrt{n-3}}\right) \approx \frac{1}{\sqrt{n-3}}$ when $n > 21$. As we convert to a normally distributed variable z, we suggest a sample size n of at least 30.

We thus categorize items using number of questions (K) and assessment sample size (n) in the following way:

- **Poor**: $PBC \leq \frac{1}{\sqrt{n-3}}$ These items at best have a confidence interval centered at 0 and thus do not have adequate predictive power for overall assessment score.
- **Acceptable**: $\frac{1}{\sqrt{n-3}} < PBC < \frac{1}{\sqrt{K}}$ These items have acceptable predictive power but may be improved with revision.
- **Good:** $PBC \geq \frac{1}{\sqrt{K}}$ These items are at least as predictive as under an "ideal" assessment where each question is independent and explains $\frac{1}{K}$ variance of student responses.

We note these categories align with common thresholds provided in the literature. For example, Varma (2006) notes "A point-biserial value of at least 0.15 is recommended, though our experience has shown that "good" items have point-biserials above 0.2 (p. 6)" which aligns with $n = 45$ for the minimum recommended PBC and $K = 24$ for the minimum good PBC. These thresholds also assume that $K < n - 3$, which may not hold for assessments administered in

small classes[2]. Considerations for thresholds where $K \geq n - 3$ will be evaluated in future studies.

Reliability

Assessment reliability refers to the correlation between a student's performance on a given assessment and their performance on a *parallel* assessment. In other words, how well does a student's performance on a given exam match their performance on a nearly identical exam, such as different versions of the same exam. The most straightforward way to compute this statistic would be to compute the correlation between a student's scores on two exam versions taken in a short time period from one another. This measure is impractical to compute in a typical learning environment. A practical measure is the Kuder-Richardson Formula 20 (KR-20) reliability coefficient (Salvucci et al., 1997). The KR-20 coefficient given by Equation 2 measures the correlation between the variance in student scores on a given exam to the proportion of students that got each question correct.

$$r = \left(\frac{K-1}{K}\right)\left(\frac{\sigma_x^2 - \sum_{i=1}^{K} p_i(1-p_i)}{\sigma_x^2}\right), \quad \sigma_x^2 = \frac{\sum_{j=1}^{n}\left(x_j - \bar{x}\right)^2}{n-1} \tag{2}$$

Where K is the number of questions on the exam being assessed, p_i is the proportion of correct answers on the i-th question, n is the number of students that completed the exam being assessed, x_j is the j-th student's score on the exam in question, $\bar{x}$ is the mean score on the exam being assessed.

(Salvucci et al., 1997) suggested the following criteria:

- **Low reliability**: $r < 0.5$.
- **Moderate reliability**: $0.5 \leq r < 0.8$.
- **High reliability**: $0.8 \leq r$. (p. 115).

Item Response Theory

In contrast to Classical Test Theory, Item Response Theory (IRT) methodologies seek to learn a student's latent (unobservable) ability variable by analyzing the students' responses to each given item in the context of all students' responses to each given item. In general, this approach works by developing a mathematical model that relates a given item's estimated difficulty (based on all students' responses) to the learned latent ability variable (based on the individual student's responses to all items). The developed model can then be used to predict the probability that a given student (with corresponding latent ability) will correctly answer a given item (with corresponding estimated difficulty). In this model, students with an equal or higher latent ability than an item's estimated difficulty are predicted to correctly answer the item.

[2] A sample size of at least n = 30 is recommended for computing PBC. This will help maintain the assumptions with regard to the normally distributed variable z used to derived the lower bound.

Rasch One-Parameter Model

In the dichotomous setting, one can construct a one-parameter Rasch model that predicts that a student will answer an item correctly using the following probability distribution function:

$$p = P(X = 1|\theta, \delta) = \frac{exp(\theta-\delta)}{1+exp(\theta-\delta)} \quad (3)$$

Where X is the binary random variable that is equal to 1 if a student answers a question correctly and 0 if they answer it incorrectly, θ is the learned latent student ability variable, and δ is the estimated difficulty of a given item. The one-parameter Rasch model assumes all items discriminate equally and thus the discrimination parameter, which would multiply $(\theta - \delta)$ in a one-parameter model, is 1. Given θ and δ for a particular student and item, respectively, Equation 3 returns the probability that the student in question answers the item correctly. Note when student's latent ability θ is equal to item difficulty δ, the probability of the student answering the question correctly is 0.5.

When we fix some item difficulty, we can plot the sigmoid curve relating ability to probability of answering the item correctly referred to as Item Characteristic Curve (ICC). Figure 1 provides ICCs for items of difficulties -1, -0.5, 0, 0.5, and 1 to illustrate how items with lower item difficulty reach higher correct probability more quickly. For example, a student with 0 estimated ability would be expected to answer an item with -1 difficulty correctly about 73% of the time while the same student would be expected to answer an item with 0.5 difficulty about 27% of the time.

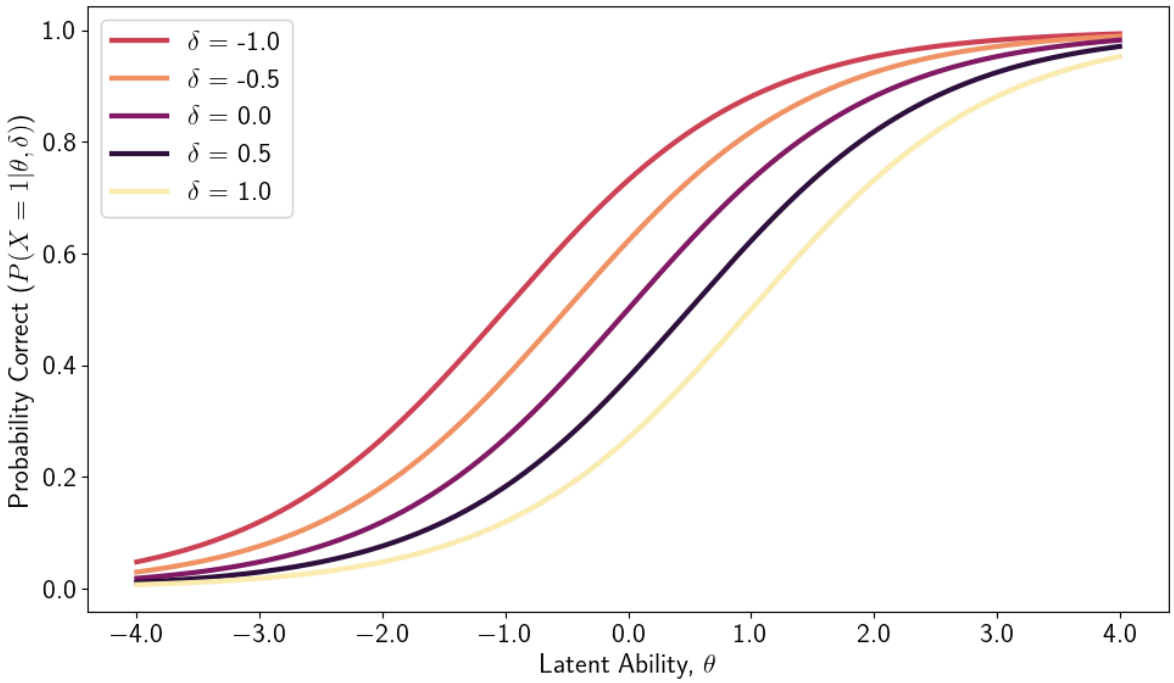

Figure 1: Item Characteristic Curve for item difficulties (δ) of -1, -0.5, 0, 0.5, and 1.

Estimated Item Difficulty and Latent Ability

Two immediate consequences of training a one parameter Rasch model are "learning" an estimate for the item difficulty for each assessment item, and the latent ability for each student that took the exam. This means, for each item and student, there is a corresponding probability distribution function from which the probability that they will get an item correct. This probability can be obtained by substituting the appropriate θ and δ into Equation 3.

Beyond the utility in computing the probability a given student will answer a question correctly (which will provide the basis for analyses to come), these derived measures from training the model can provide insight into the student population and the questions on the assessment. The distribution of student abilities and item difficulties can help inform the assessor of the spread of the students' abilities and the items' difficulty. It is important to note that the latent ability θ and item difficulty δ are unitless, and do not represent the same type of numerical relationship that PBC or item difficulty represent.

Infit and Outfit

After constructing a Rasch model of estimated student ability and item difficulty to predict the probability of answering a question correctly, one can consider how well the model fits the data. The Rasch model makes the assumption that all items have the same discrimination parameter value 1. We can consider fit values of less than 1 as overfitting the data and fit values greater than 1 as underfitting. We consider two residual-based fit statistics: outfit (outlier-sensitivity fit) and infit (information-weighted fit).

Residual-Based Fit

Residuals, in the case of a Rasch model, refers to the difference in the model's calculated probability of a student, s, answering an item, i, correctly: $E(x_{si}) = P(X = 1|\theta_s, \delta_i)$, and the student's observed score, x_{si}, on the same question. The residual-based fit is calculated as this difference divided by the variance of the item response.

$$z_{si} = \frac{x_{si} - E(X_{si})}{Var(X_{si})} \tag{4}$$

Residual-based fit statistics can be measured against the standardized error of items as defined by 1 over the sum by items of the fit variance.

$$SE_i = \frac{1}{\sum_i v_{si}} \tag{5}$$

Outliers occur beyond $1 \pm SE_i$. Since items with fit near 1 have the least discrimination power, it is suggested to use items with fit less than 1. We can thus categorize fit statistics as follows:

- **Poor:** $fit \geq 1 + SE_i$
- **Acceptable:** $1 - SE_i < fit < 1 + SE_i$

- **Good:** $fit \leq 1 - SE_i$

Outfit (outlier-sensitivity fit)

The outfit, or unweighted mean fit, is the average of the standardized fits without accounting for item variance.

$$outfit = \frac{\sum_n z_{si}^2}{N} \tag{6}$$

Outfit identifies patterns when item difficulty is far from student ability. For example, outfit might identify when students with high ability answer a low difficulty question incorrectly or vice versa.

Infit (information-weighted fit)

The infit, or weighted mean fit, is the average of the standardized fits while accounting for item variance.

$$infit = \frac{\sum_n z_{si}^2 Var(X_{si})}{\sum_n Var(X_{si})} \tag{7}$$

Infit identifies patterns when item difficulty is close to student ability. For example, infit might identify when students with ability near the item difficulty perform far better than expected by the model.

Distractors

Plausible, but incorrect, answers to the problem are referred to as *distractors*. Distractors can represent a wide variety of types of incorrect responses, and have been the topic of plenty of literature (Chamberlain Jr. & Jeter, 2019, 2020; Gierl et al., 2015, 2017). They can represent student misconceptions about the concept being assessed in a question or another concept; they can represent issues in manipulation or representation; they can even be created by manipulating the correct answer to the problem. Given the many types of thinking distractor answer choices can represent, they potentially play an integral role in the discriminative power of a multiple choice item. For example, an item that has no distractors that represent plausible misconceptions could inadvertently direct students to the correct answer choice without any understanding of the underlying concept being evaluated.

Distractors Chosen Percentage

In light of the significance of distractor answer choices, the frequency with which distractors are chosen can be very telling. The percentage of chosen distractors measures the rate at which students chose a distractor against the total number of distractors on the exam. We categorize

distractors into three categories based on how often they were selected: “never chosen” (chosen 0% of the time), “rarely chosen” (chosen between 0% and 5% of the time) and “sometimes chosen” (chosen more than 5% of the time). These categories help to identify distractors that were poor (never chosen), distractors that were either chosen at random or elicited a small percentage of students (potentially effective distractors), and effective distractors. Distractors in the latter two categories have potential to give insight into misconceptions present in student thinking, while distractors in the first category should be discarded and replaced with distractors based on theoretical or experimental design.

Effective Distractors

An effective distractor is an incorrect answer choice that is chosen by students at a rate of at least 5% (Hingorjo & Jaleel, 2012). A distractor choice being chosen with some level of frequency indicates that it is likely to capture potential coherent thinking resulting from misconceptions about content. Ideally, every distractor would be “effective,” (each question would have 3 or 4 effective distractors, depending on whether the question has 4 or 5 options) meaning that it is chosen at least 5% of the time. Realistically, many factors contribute to the choice of a distractor on an item, such as the exam it is seen on (e.g., first exam versus final exam) and the level of students taking the exam. Moreover, the literature suggests that 3 choices (2 distractors and 1 correct solution), is optimal from a theoretical, empirical, and practical consideration (Haladyna et al., 2019). In light of these potential issues, the number of effective distractors a question has should hint to the discriminative power of the question itself.

Identifying Discriminative Items

Utilizing item measures from Classic Test Theory, Item Response Theory, and Distractor Analysis, we present a summary of each metric that will inform labeling items as sufficiently discriminating for the purposes of an educational assessment. Our methodology begins by classifying each item-level metric into “ideal”, “acceptable”, or “poor” and analyzing how these categories collectively explain the discriminative power of the item.

Item-Level Metrics

- **Item Difficulty** - Proportion of student sample who answered the item correctly. Ideally, item difficulty is normally distributed about 0.74 with as small a variance as possible. We consider items within one standard deviation of 0.74 to be “ideal” and items within two standard deviations as “acceptable.” Items outside of this range are “poor.”
- **Point-Biserial Correlation** - Correlation between proportion of student sample who answered the item correctly and student sample overall score on assessment. Ideally correlation is at or above $1/\sqrt{K}$ and is at least above $1/\sqrt{n-3}$.
- **Rasch Item Infit** - Infit identifies patterns when estimated item difficulty is close to estimated student ability. Ideally less than 1 - standard error, at least less than 1 + standard error.
- **Rasch Item Outfit -** Outfit identifies patterns when estimated item difficulty is far from estimated student ability. Ideally less than 1 - standard error, at least less than 1 + standard error.

- **Effective Distractors** - Distractors chosen at least 5%. Ideally all non-answers are effective Distractors, while at least one non-answer is an effective distractor.

We then classify each assessment-level metric into “ideal”, “acceptable”, or “poor” and analyze how these categories collectively explain the reliability and discriminative power of the assessment.

Assessment-Level Metrics

- **Item Difficulty Distribution** - Ideal average item difficulty is slightly higher than halfway between random guess chance and 100% (e.g., 4-option item as ideally 0.74) with minimal item difficulty variation.
- **KR-20 Reliability Coefficient** - Reliability between different forms of an assessment. Ideally at or above 0.8, at least at or above 0.5.
- **Distractors Chosen Distribution** - Distribution of Distractors chosen across all items. Ideally Distractors are chosen at least 5% of the time.

Methods

To present a working example of our methodology, we study four multiple-choice exams administered in College Algebra over the course of the Fall 2017 term at the University of Florida. The course consisted of 240 students divided over 9 sections. Table 1 presents a summary of the assessments.

		Exam 1		**Exam 2**			**Exam 3**			**Exam 4**		
		A	B	A	B	C	A	B	C	A	B	C
Number of MC Questions	3 options			1	1	1						
	4 options	14	14	15	15	15	17	17	17	25	25	25
	5 options	2	2									
	total	16	16	16	16	16	17	17	17	25	25	25
Students that took the exam		117	123	75	76	77	75	72	73	76	75	74

Table 1 Summary of the multiple-choice components of the exams administered in Fall 2017.

To perform a thorough analysis of the exams, the exam items, and the students, we perform both exam-level analyses using Classical Test Theory (CTT) and item-level analyses using Item Response Theory (IRT).

Classical Test Theory

The CTT methods outlined in the methodology require extracting every student answer to each exam item as well as the corresponding solution to each item. With this information, observed scores for each student on each assessment can be computed by dividing the number of items they answered correctly on an exam by the number of items on the exam. Similarly, item difficulty is computed by dividing the number of correct responses to an item by the number of students that attempted the item. With the observed score for each student and their responses to each item, point-biserial correlation can then be computed by evaluating Equation 1 for each

item and computing the corresponding bounds based on the number of questions on an exam and the number of students that took that exam. Lastly for CTT analyses, using the responses from each item (and in turn the proportion of correct answers), KR-20 can be computed for each exam directly from Equation 2.

Item Response Theory

We construct a one parameter Rasch model using the unconditional maximum likelihood estimation (UCON) method. The steps for training the model are as follows:

1. **Remove students with 100% scores and 0% scores**
   As we are computing the natural log of average scores (either by student or by item), log(0) will produce errors.
2. **Approximate ability and difficulty**
   a. **Estimate ability per student** using: $\theta_s = \ln\left(\frac{\delta_s}{1-\delta_s}\right)$, where $\delta_s = \frac{\sum q_i}{n}$ *(e.g., sum of 1/0s for a single student divided by number of items student answered)*
   b. **Estimate difficulty per item** using: $b_i = \ln\left(\frac{1-\mu_i}{\mu_i}\right)$, where $\mu_i = \frac{\sum q_s}{n}$ *(e.g., sum of 1/0s by all students for a single question divided by number of students who answered the question)*
3. **Standardize difficulty for mean 0**
   Adjust difficulty per item with: $\beta_i = b_i - \bar{b}$ *(e.g., adjusted difficulty is estimated difficulty subtracted by average difficulty of all items)*
4. **Iterate until sum of squares of residuals is sufficiently close to 0**
   a. ***Calculate expected values:*** probability of student s answering question i correctly given a student's ability score and the item's difficulty $P(\theta_s, \beta_i)$:

$$P(\theta_s, \beta_i) = \frac{e^{\theta_s - \beta_i}}{1 + e^{\theta_s - \beta_i}}$$

   b. ***Re-calculate all*** θ_s ***and*** β_i using the new expected values
   c. ***Calculate estimated variances of expected values:*** $v_{si} = P(x_{si}) * (1 - P(x_{si}))$
   d. ***Calculate residuals between estimates and original data***: $e_{si} = q_{si} - P(x_{si})$
   e. ***Calculate sum of squares of residuals***:

$$e = \sum_i \sum_s \left(e_{si}\right)^2$$

   f. **If e > 0.0001, re-calculate expected values and repeat:** Calculate revised θ_s and β_i and restart at step 4 a.

$$\theta_s = \theta_s + \frac{\sum_s e_{si}}{\sum_s v_{si}} \qquad \beta_i = \beta_i - \frac{\sum_i e_{si}}{\sum_i v_{si}}$$

Distractors

The distractor-based methods outlined in the Methodology only require compiling a list of all student answers and a complete list of possible answer choices with their label of "distractor" or "solution." From these data, percentages of distractors chosen and effective distractors per item only require direct computations of each distractor's frequency of being chosen and then counting the number of distractors per item that were chosen more than 5% of the time.

Item Discrimination

A summary dataset of all of the items can be constructed using the methods for generating item level measurements outlined in the previous "Methods" subsections. With this summary dataset, the "ideal/good," "acceptable," and poor thresholds can be directly computed, and each item directly assessed for its quality with regard to each measure.

Results

Here we present an overview of each measurement described in the methodology with respect to the exams outlined in the Methods section.

Classical Test Theory

Overall, the Classical Test Theory analysis paints a picture typical of multiple-choice assessments in lower division Mathematics courses: there is a tendency for a large percentage of the class to get nearly all the assessment items correct. Ultimately, this limits the amount of discriminative power of assessment items to determine specific conceptual misunderstandings, because the modal response to a given item is correct.

Observed score distribution

Figure 2 presents an overview of the observed scores across the 11 exam/form combinations. While both forms for exam 1 had observed scores that were roughly normally distributed about 70%, the remaining exams are more closely modeled by a beta distribution, given the significant mass on the right-hand side of the distribution (all distributions having a fairly large negative skewness). This is in contrast for the preferred observed score distributions described in the Methodology. Ideally, observed scores should follow a normal distribution with a kurtosis (the measure of the probability mass contained in the tails of the distribution) close to zero.

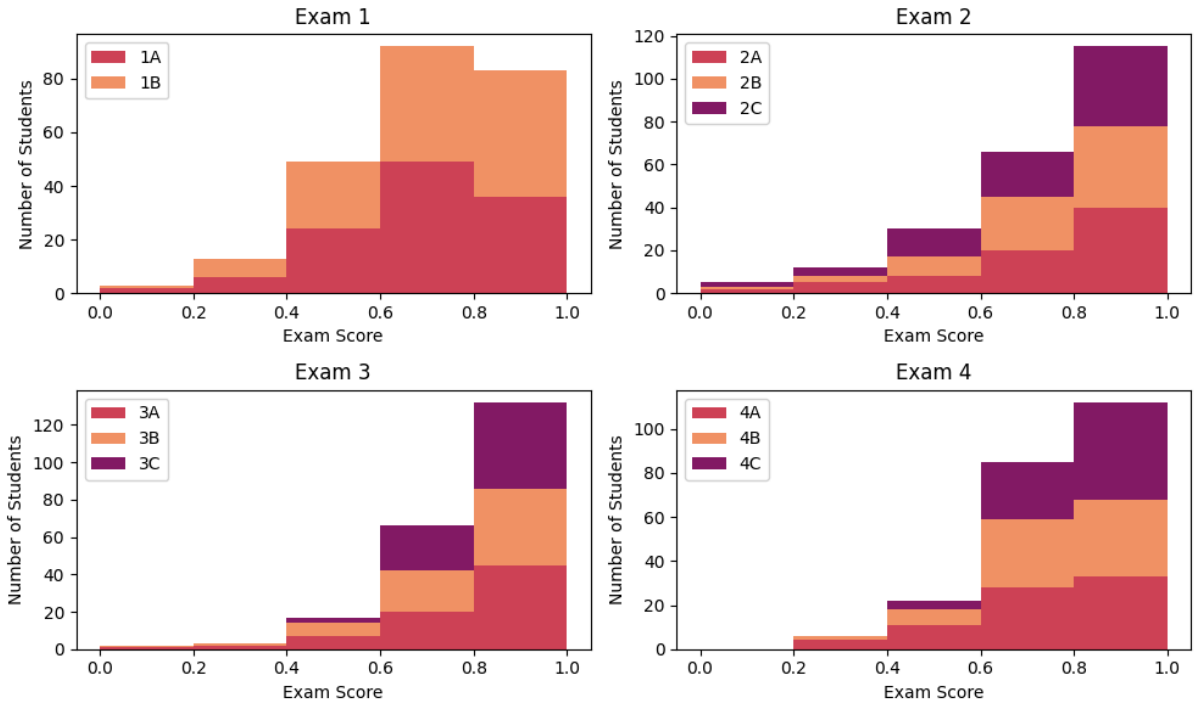

Figure 2 Histograms showing the distribution of observed scores as a percentage for each student and each exam.

The tendency towards right-skewed distributions can limit the discriminating power of the exams. An exam composed almost entirely of items that most of the students get correct will not be able to adequately discriminate between students with and without conceptual understanding. Alternatively, it suggests that questions can be dropped without affecting how well the exam can discriminate between student understanding.

Reliability

We measure assessment reliability using the Kuder-Richardson Formula 20 (KR-20) correlation coefficient. It measures the correlation between the proportion of right and wrong answers among all questions on an exam and provides a metric for assessing the internal consistency of an exam of dichotomous questions.

As we detailed above, a KR-20 below 0.5 means an exam has low reliability, KR-20 between 0.5 and 0.8 means an exam has moderate reliability, and KR-20 > 0.8 has high reliability. The KR-20 values for the 11 exams are presented in Table 2.

Table 2 KR-20 correlation coefficients for the 11 exams studied.

Exam	**Exam 1**		**Exam 2**			**Exam 3**			**Exam 4**		
Form	A	B	A	B	C	A	B	C	A	B	C
KR-20	0.68	0.71	0.78	0.71	0.74	0.78	0.75	0.53	0.77	0.77	0.70

Aside from Exam 3 Form C, all of the exams scored comfortably in the "moderate reliability" range. This means that a student given a parallel, but non-identical exam should achieve a similar observed score to the exam that they completed during the period of study.

Item difficulty

Item difficulty describes the proportion of students that answer a given dichotomous question correctly. For dichotomous exams, it is ideal for item difficulty to be normally distributed with a small variance, such that there is sufficient data to discriminate between high and low performing students.

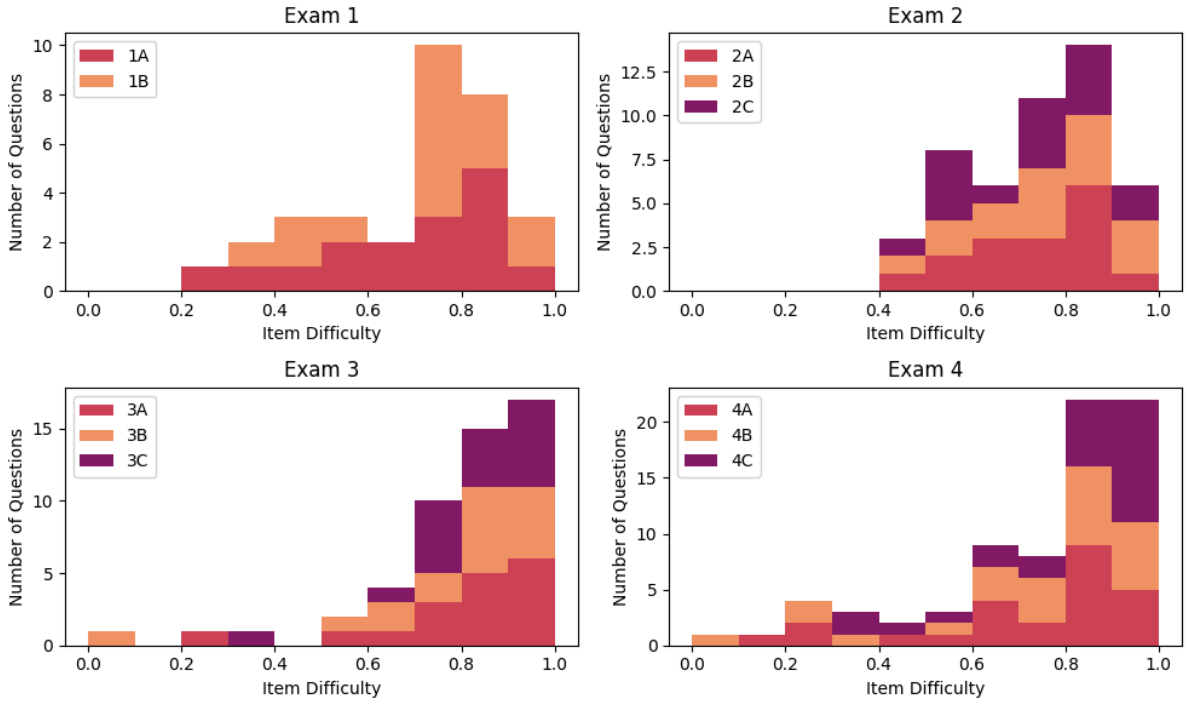

Figure 3 Histograms showing the number of items for a given difficulty bin for each of the 11 exams.

Figure 3 presents the distribution of item difficulty across the 11 exams in the study.

Similar to the observed score distributions, with the exception of exam 1, the distributions for item difficulty are not normal, they are mostly right skewed. Combined with the observed scores distributions, this indicates that not only do most students perform well on all of the exams, but most of the items are answered correctly by most students.

Point-biserial correlation

Using the formula in Equation 2 and the thresholds for "poor," ($PBC \leq \frac{1}{\sqrt{n-3}}$) "acceptable," ($\frac{1}{\sqrt{n-3}} < PBC < \frac{1}{\sqrt{K}}$) and "good" ($PBC \geq \frac{1}{\sqrt{K}}$), we are able to classify the point-biserial

correlation for each item on each exam. The number of questions in each category for each exam is presented in Figure 4.

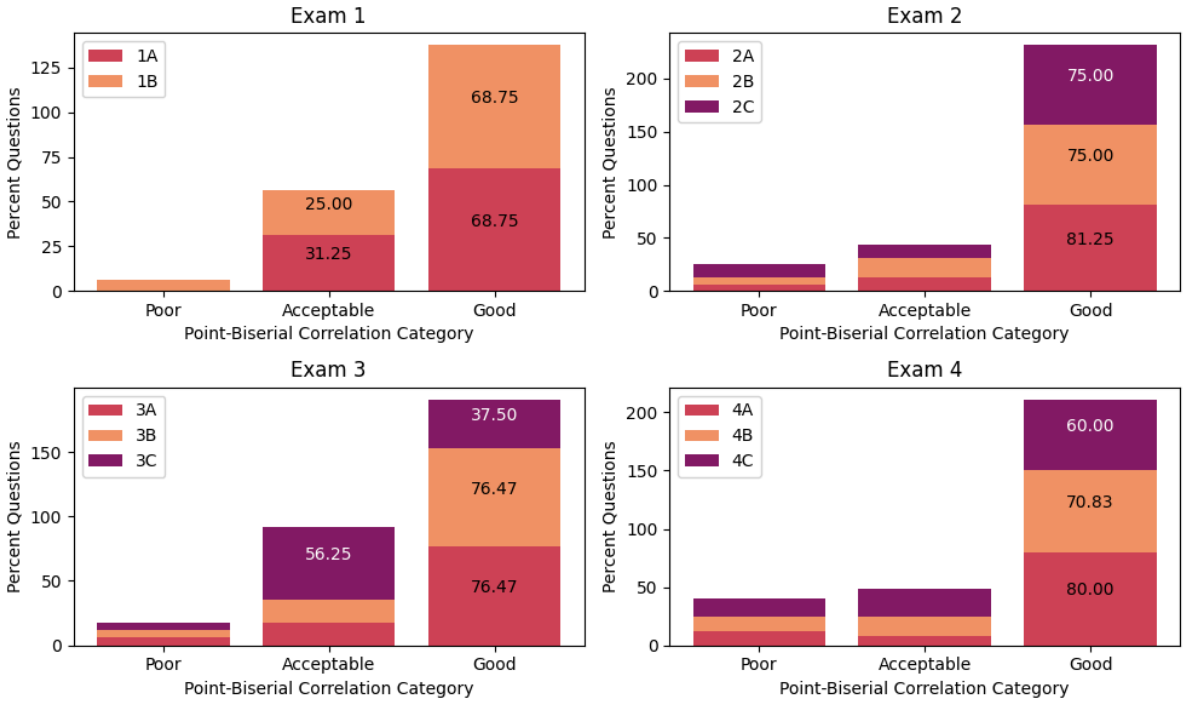

Figure 4 Bar chart showing the percent of questions on each exam that have a "poor," "acceptable," and "good" point-biserial correlation. Percentages for bars of the same color add up to 100%.

No exam has more than a few “poor” items in terms of point-biserial correlation, in fact, most questions on most exams fall into the “good” category. This indicates that for most questions, students’ performance on that question is indicative of their overall performance on the exam. This is consistent with the general tendency towards most students answering most questions correctly.

Item Response Theory

Overall, IRT results suggest exams 1 and 2 provide few outliers and that all questions discriminate student ability in some manner. However, exams 3 and 4 produced several outliers in both low and high difficulty questions evident through the estimated item difficulty values outside of -2 to 2 range, poor/good infit, and poor/good outfit. We present a more detailed look at each exam and each metric below.

Estimated Item Difficulty and Latent Ability

Item difficulty for exams 1 and 2 follow a normal distribution with almost all items estimated between -2 and 2. However, exams 3 and 4 present a small number of questions outside of the range -2 to 2 which are likely to be outliers. Moverover, exam 4 in particular presents a

skew-right distribution of item difficulty with a peak around -1, suggesting this exam was easier in a sense than the other 3. Item difficulty histograms for exams 1-4 are provided in Figure 5.

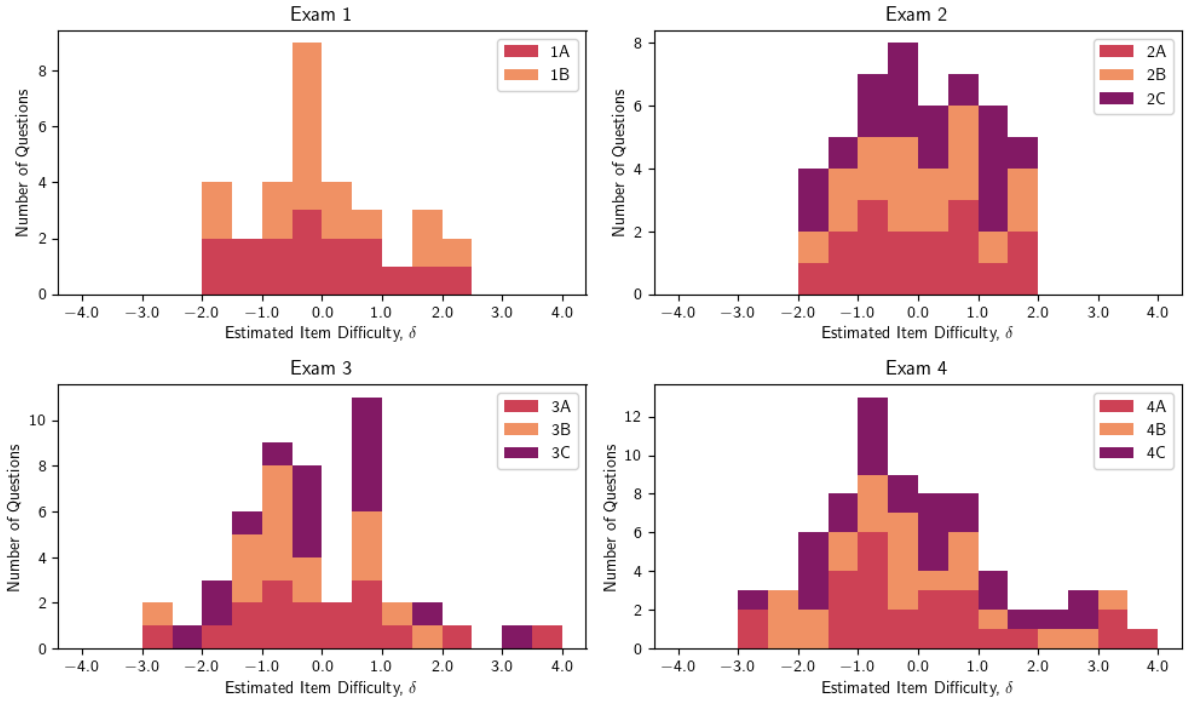

Figure 5 Distributions of the estimated item difficulty δ for each exam learned from training the one parameter Rasch model described in Equation 3.

Student estimated latent ability over all 4 exams appeared skew-left with a peak between 1 and 2. Recall that estimated latent ability begins as the natural log of a ratio of each student's exam scores. The model then adds to student ability to better fit the sigmoid curve and distinguish between student performance. The model also removes any student scores of 0 and 100, resulting in the low number of highest estimated latent ability.

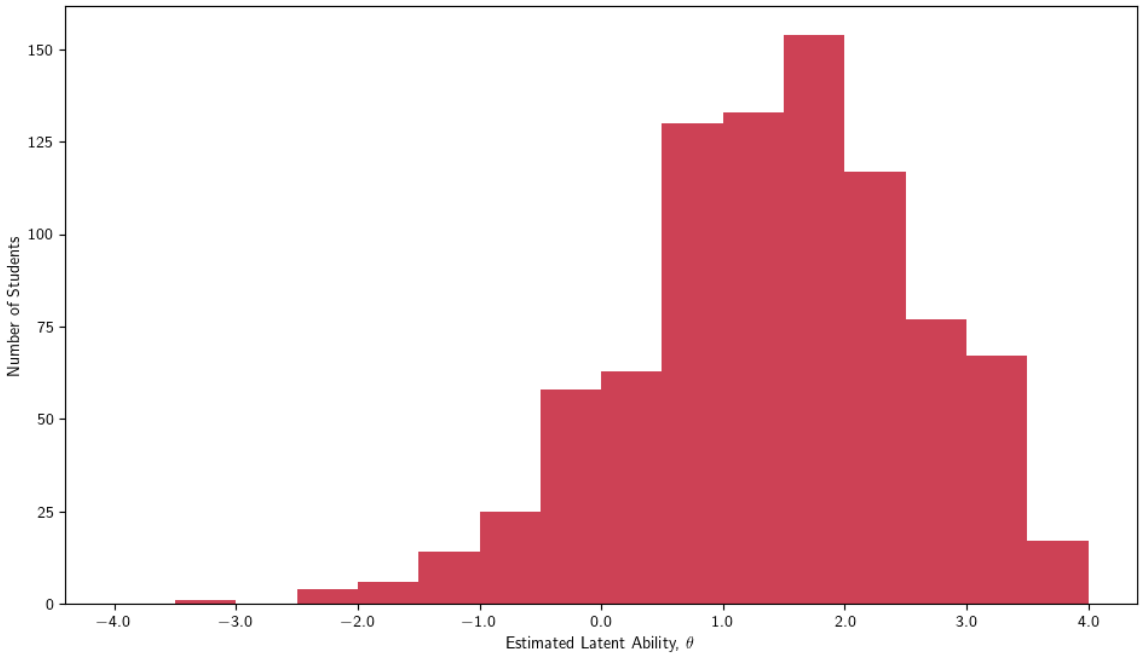

Figure 6 Distributions of the estimated latent ability θ *for each student learned from training the one parameter Rasch model described in Equation 3.*

Item Infit

Recall that Item Infit identifies patterns when item difficulty is close to student ability. Item Infit for each exam are presented in Figure 7 and shows that the Rasch model fits the large majority of all exam items. It is unlikely that infit alone will predict the discrimination of any item, though a poor item infit may be particularly indicative of an item that needs to be considered for removal.

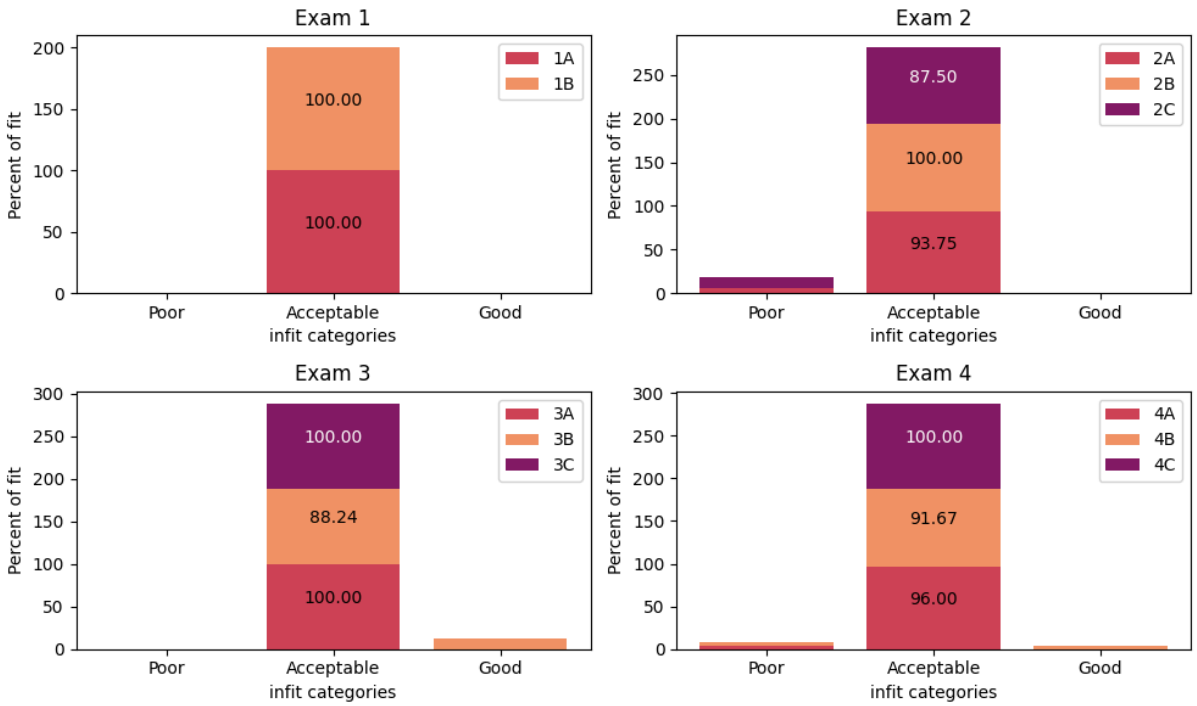

Figure 7 Distributions of the infit categories for each exam described in the "Infit" section of the "Methodology."

Item Outfit

Recall that outfit identifies patterns when item difficulty is far from student ability. Unlike Item Infit, Figure 8 illustrates that all four exams show some good outfit (overfit outliers) and poor outfit (underfit outliers). As expected with the increased frequency of item difficulties outside of -2 and 2 in exams 3 and 4, far more outfits of poor and good appeared in those exams.

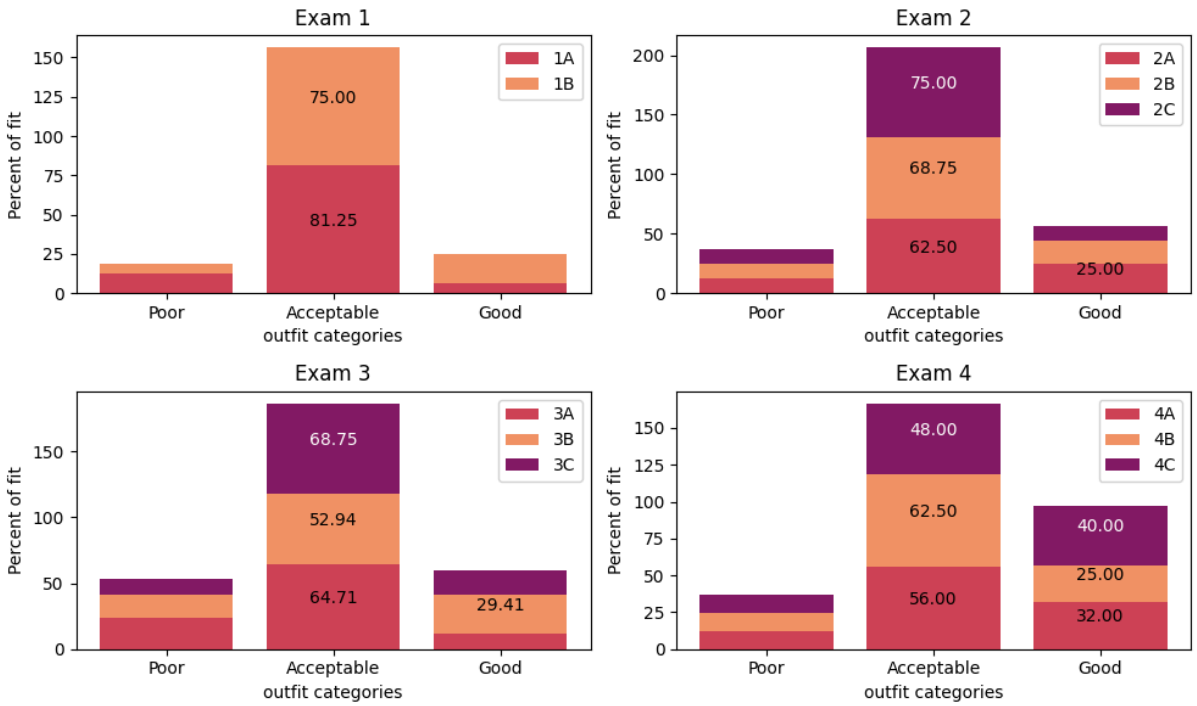

Figure 8 Distributions of the outfit categories for each exam described in the "Outfit" section of the "Methodology."

To further investigate the differences between poor and good item outfit, we consider the correlation between good/poor outfit and estimated item difficulty as well as good/poor outfit and point-biserial correlation. Figure 9 presents the results of graphing these 4 pairs and presents both the normal and natural log correlation for each.

At a glance, both good and poor outfit are correlated with the items' estimated difficulties. In particular, good outfit occurred primarily in questions with difficulties between -2 and 1 while poor outfit was distributed throughout item difficulty (between -2 and 4). In other words, "easier" questions had the best fit with the model (smallest difference between expected probability of success and actual success). On the other hand, good and poor outfit do not appear correlated with point-biserial correlation. This suggests outfit and point-biserial correlation can be used in tandem to further distinguish between good and poor items on an assessment.

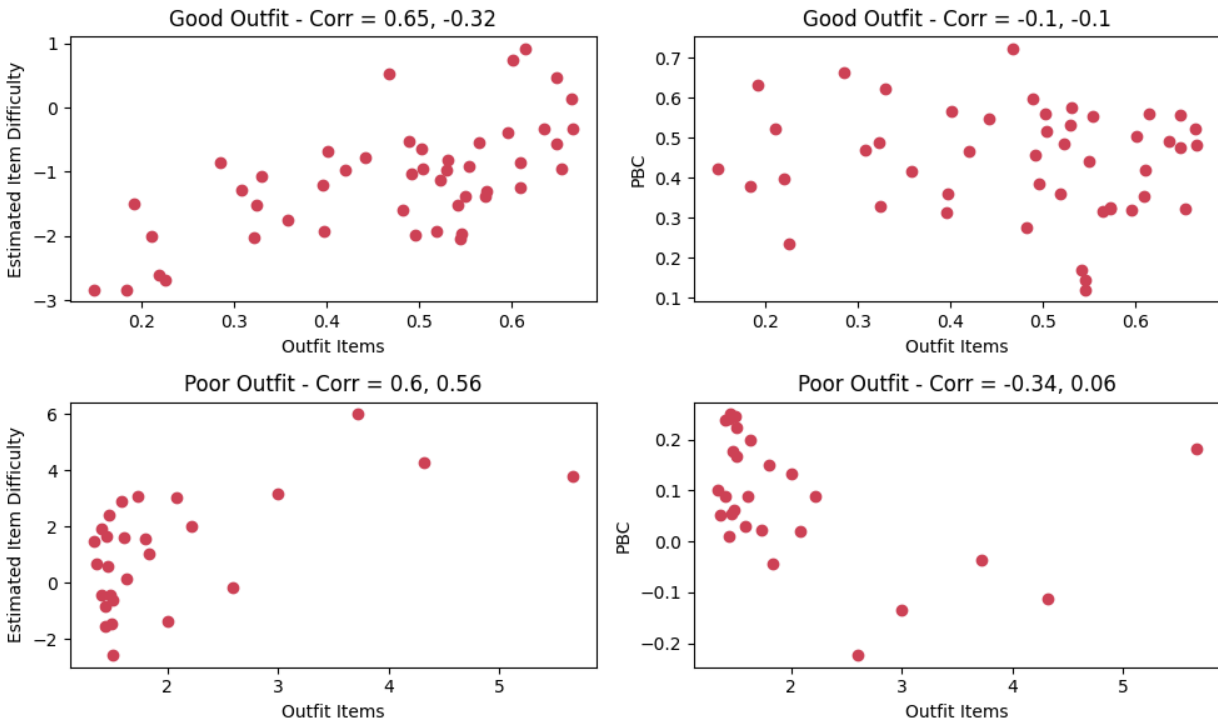

Figure 9 (top left) Scatterplot comparing item outfit and estimated item difficulty for items with "ideal/good" outfit. (top right) Scatterplot comparing item outfit and point-biserial correlation (PBC) for items with "ideal/good" outfit. (bottom left) Scatterplot comparing item outfit and estimated item difficulty for items with "poor" outfit. (bottom right) Scatterplot comparing item outfit and point-biserial correlation (PBC) for items with "poor" outfit.

Distractors

Presenting answer choices that effectively capture students' misconceptions of the content mastery being evaluated can markedly improve the diagnostic utility of summative assessments. In turn, this targeted distractor creation methodology can improve the discriminative power of assessment items by providing an appropriate answer choice for students with specific misconceptions (leading them to select the distractor answer choice instead of potentially selecting at random). Similarly, poorly constructed distractor choices can increase the probability of selecting the correct answer choice for savvy students by allowing them to eliminate answer choices without any content mastery.

The results from analyzing the distractors in our study point to many distractors across the item set being selected less than 5% of the time. Subsequently, the percent of items with more than one effective distractor (a distractor chosen more than 5% of the time) is around 50% for most of the exams. This implies that on most of the items, students are able to effectively reduce the number of likely answer choices to 2 (the solution and the effective distractor). This increase in probability of answering an item correctly at random (from 25% to 50% for most questions) dramatically limits the discriminative power of these items.

Distractors Chosen Percentage

Analyzing the frequency with which a distractor answer choice is chosen, the "distractors chosen percentage", allows one to gain insight into how well the question design overall is able to capture misconceptions into student thinking. Distractor answer choices can be split into three major categories depending on the frequency with which they are chosen: chosen 0% of the time ("never chosen"), chosen 0–5% of the time ("rarely chosen"), and chosen > 5% of the time ("sometimes chosen").

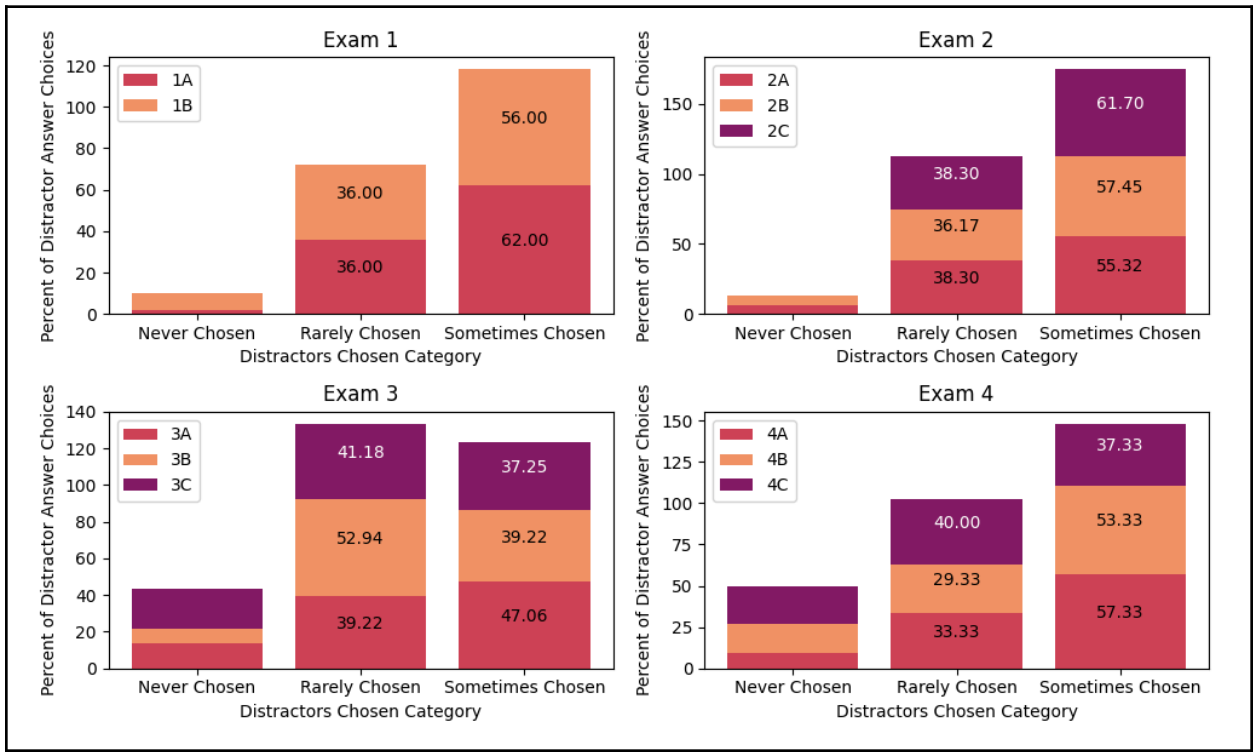

Figure 10 Bar chart showing the percent distractor answer choices chosen for the following categories: "Never Chosen" (chosen 0% of the time), "Rarely Chosen" (chosen between 0% and 5% of the time), "Sometimes Chosen" (chosen more than 5% of the time). Percents in each color of blocks sum to 100%.

Figure 10 presents the percentage of distractors chosen in each category described above. It is noteworthy that each exam has between 37% and 62% of distractors chosen more than 5% of the time. This indicates that largely distractor answer choices are scarcely selected. This mirrors the findings related to observed scores: most students selected the correct answer on most items most of the time.

Effective Distractors

With the cutoff of being selected at least 5% of the time for a distractor to be considered "effective," Figure 11 shows the percent of questions with a given number of effective distractors for each exam. Most exams have $\approx$ 50% of questions with one or fewer effective distractors. This means that there are a significant number of questions (though not all of the 50%) that a student was able to improve their probability of randomly selecting the correct answer to at least 50%.

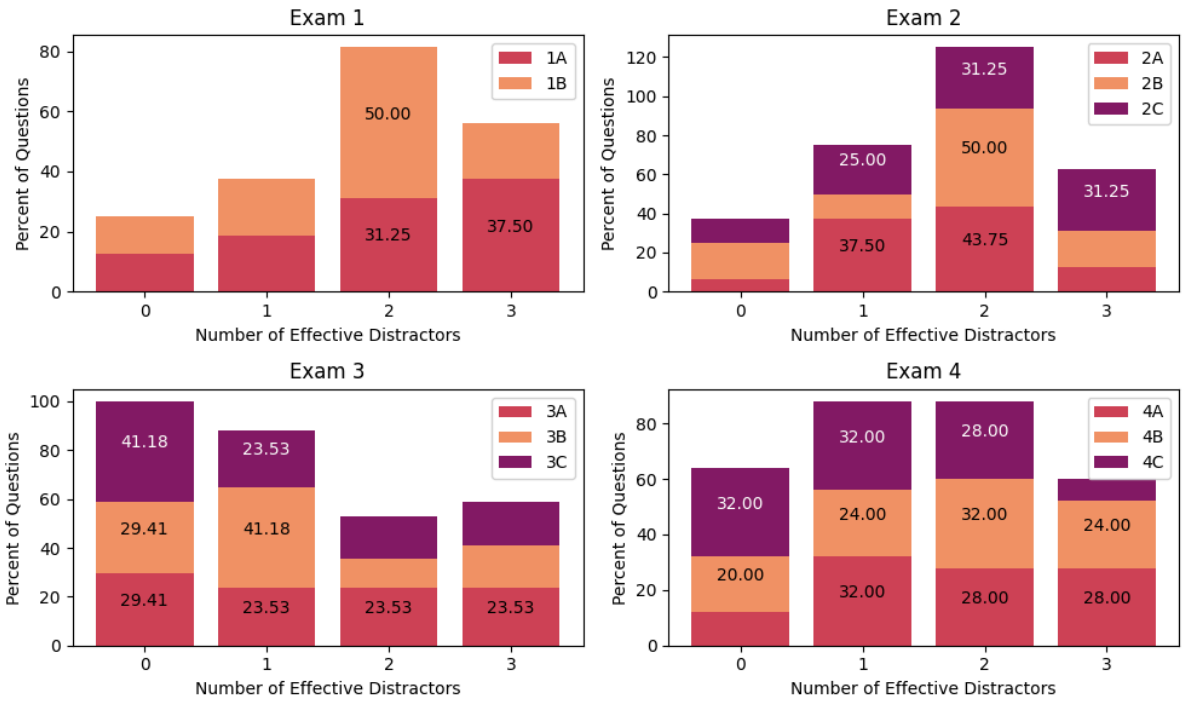

Figure 11 Bar chart showing the percent of questions with 0, 1, 2, and 3 effective distractors. Percents in each color of blocks sum to 100%.

Identifying Discriminative Items

Combining all of the measurements for assessing assessment items with their corresponding "ideal/good," "acceptable," and "poor" categorizations results in the heat map shown in Figure 12. Each item in exam 4B is displayed with the six key measurements for identifying the discriminative power displayed on the x axis: estimated item difficulty, outfit, infit, number of effective distractors, point-biserial correlation, and item difficulty. The cells for each item are color coded depending on whether the measurement falls into the "ideal/good," (green) "acceptable," (yellow) or "poor" (red) category.

The first takeaway from Figure 12 is the need to consider more than CTT or IRT metrics in isolation. Item 4B26 has an estimated item difficulty of 4.28, which should be highly discriminative, but the PBC of -0.11 and item difficulty of 0.09 suggest that this item was so hard that almost no students were able to answer it correctly. While this item may be able to identify the top student in a class, it is not generally effective at discriminating between high and low performing students because it is so hard.

Items 4B09 and 4B10 are illustrative examples for the efficacy of the integrated methodology. According to Rasch estimated item difficulty, infit, and outfit, both 4B09 and 4B10 are good candidates for discriminative items. However, combined with effective distractors, PBC, and item difficulty, it is clear that only 4B10 is an effective item. 4B09 has no effective distractors, and is too easy to be discriminative (96% of students answered it correctly).

Lastly, the methodology can quite effectively identify ineffective items. 4B19 falls into the "poor" category for every metric considered except for effective distractors and item difficulty (both are acceptable).

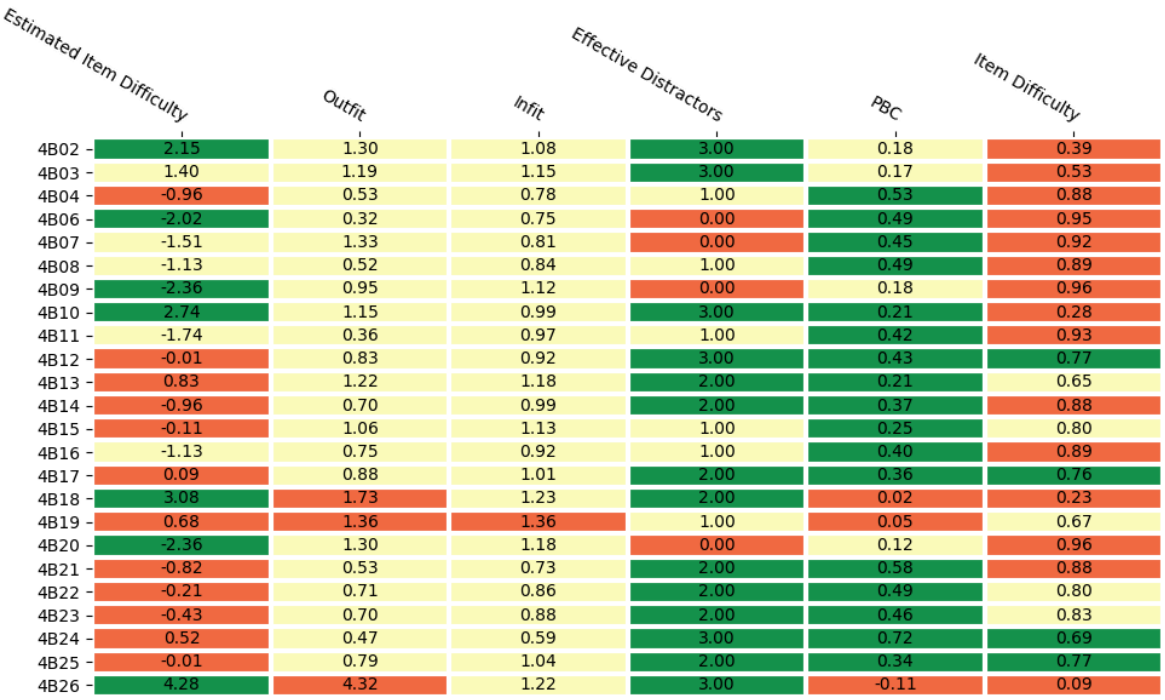

Figure 12 Heat map showing the item summary for exam 4B. Green items achieve the "ideal/good" thresholds outlined in the Methodology section, yellow items achieve the "acceptable" thresholds, and red items are "poor."

Discussions

The integrated methodology resulting in Figure 12 shows how insufficient CTT and IRT analyses are in isolation, but how effective they can be when combined. The thresholds provided with the integrated methodology provide a direct comparison strategy that allows an instructor evaluating the items in an assessment convenient bounds for judging the discriminative power of each item across a range of measurements. When the measurements are in agreement, this provides the instructor confidence in the discriminative power of an item and will help them decide whether or not to maintain that item for future assessments. Moreover, the inclusion of distractor metrics furthers the educational purposes of assessments by identifying patterns of student thinking rather than solely predicting final score on assessment. It also provides nuance to considerations for the optimal number of options on an item. Rather than arguing from a theoretical perspective that 3 options provide the best statistical predictiveness, the number of item options may increase if more than 2 effective distractors can be established on an item.

Returning to the goals of different types of assessments, these metrics may be used to further specific goals of an assessment. For example, summative assessments will pay closer attention to item difficulty and PBC as the goal of the assessment is to provide a single "descriptive" score associated with the student's ability. Alternatively, formative assessments like assessments as learning (AaL) will pay closer attention to effective distractors (especially to present feedback associated to established patterns of student thinking) and outfit outliers as both can identify

patterns of students answering incorrectly in expected ways (distractors) and unexpected ways (outfit outliers).

Conclusions

In this study we presented an integrated methodology that combines ideas from classical test theory, item response theory, and distractor analysis to identify discriminative items in multiple choice mathematics assessments. We provided explicit, statistically motivated thresholds for each measurement so that each item could be procedurally categorized as "ideal/good," "acceptable," and "poor" with respect to each measurement. This methodology provides instructors and researchers a high level view of each item in an assessment that can be easily interpreted to see the relationships between each measurement and suggest items that are effective and ineffective at discriminating between high and low performing students. This allows the instructor to quickly evaluate items for retention and removal for future exams. Beyond the value to an instructor deciding which items are effective and should be retained, the items highlighted by this integrated methodology present the opportunity to determine why certain items are more effective than others, and should be qualitatively analyzed by researchers. Qualitative analysis of distractors identified using this integrated methodology remains to be studied and will be the subject of future work.

Authors' Contribution

RJ and DC designed and conceptualized the study and contributed to project administration. DC collected and processed the data. RJ and DC contributed to the methodology, investigation, and formal analysis. All authors interpreted the results and drafted and edited the manuscript.

Acknowledgements

DC acknowledges the support of NSF award number 2044302.

References

Chamberlain Jr., D., & Jeter, R. (2019). Leveraging cognitive theory to create large-scale learning tools. *22nd Annual Conference on Research in Undergraduate Mathematics Education*.

Chamberlain Jr., D., & Jeter, R. (2020). Creating Diagnostic Assessments: Automated Distractor Generation with Integrity. *Journal of Assessment in Higher Education*, *1*(1), Article 1. https://doi.org/10.32473/jahe.v1i1.116892

Dann, R. (2014). Assessment *as* learning: Blurring the boundaries of assessment and learning for theory, policy and practice. *Assessment in Education: Principles, Policy & Practice*, *21*(2), 149–166. https://doi.org/10.1080/0969594X.2014.898128

Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, Analyzing, and Using Distractors for Multiple-Choice Tests in Education: A Comprehensive Review. *Review of Educational Research*, *87*(6), 1082–1116. https://doi.org/10.3102/0034654317726529

Gierl, M. J., Lai, H., Hogan, J. B., & Matovinovic, D. (2015). A Method for Generating Educational Test Items that are Aligned to the Common Core State Standards. *Journal of Applied Testing Technology*, 1–18.

Haladyna, T. M., Rodriguez, M. C., & Stevens, C. (2019). Are Multiple-choice Items Too Fat?

Applied Measurement in Education, *32*(4), 350–364. https://doi.org/10.1080/08957347.2019.1660348

Hingorjo, M. R., & Jaleel, F. (2012). Analysis of one-best MCQs: The difficulty index, discrimination index and distractor efficiency. *JPMA. The Journal of the Pakistan Medical Association*, *62*(2), 142–147.

Ho, A. D., & Yu, C. C. (2015). Descriptive Statistics for Modern Test Score Distributions: Skewness, Kurtosis, Discreteness, and Ceiling Effects. *Educational and Psychological Measurement*, *75*(3), 365–388. https://doi.org/10.1177/0013164414548576

Lord, F. (1955). Estimation of Parameters from Incomplete Data. *Journal of the American Statistical Association*, *50*(271), 870–876.

Salvucci, S., Walter, E., Conley, V., Fink, S., & Mehrdad, S. (1997). *Measurement Error Studies at the National Center for Education Statistics* (Statistical Analysis Report NCES 97464). National Center for Education Statistics. https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=97464

Wu, M., Tam, H. P., & Jen, T.-H. (2016). *Educational Measurement for Applied Researchers*. Springer Singapore. https://doi.org/10.1007/978-981-10-3302-5