International Journal of Research in Mathematics Education manuscript No.
(will be inserted by the editor)

Leveraging research and technology to make diagnostic multiple-choice assessments possible

Darryl Chamberlain Jr. · Russell Jeter

Received: 6 May 2020 / Accepted: Under Review

Abstract Multiple-choice assessments are widely used for their ease of implementation and grading. Yet, these assessments are largely criticized as being unable to provide diagnostic information about student knowledge for a variety of reasons. We have proposed a method to utilize current research on student knowledge in unison with technology to make diagnostic multiple-choice assessments possible. This paper reports on the results of implementing our diagnostic multiple-choice assessments in a College Algebra course over a three year span. Quantitative and qualitative analysis suggests these assessments associate student knowledge with their choices and thus can be used as practical diagnostic tools. We conclude with how the construction of these assessments could be utilized to provide targeted feedback as students develop their understanding through non-instructor-mediated assessments such as homework.

Keywords Multiple-choice Assessment · Automated Item Generation · Distractors

1 Introduction

Assessment is a critical component of every course. While there are numerous ways to assess students' knowledge, multiple-choice tests are the most widely used assessments in K-16 as they can be the most efficient to administer while simultaneously being quick and objective to grade (Rodriguez 2011). We use a typical College Algebra item to contextualize multiple-choice item terminology in mathematics.

A multiple-choice item consists of a stem and options. The *stem* includes the context, content, and problem for the student to answer. In the example in Figure 1, this includes the instructions (context) and the problem. By *problem*, we refer to the content issue that must be solved. In Figure 1, this would be solving the linear equation. Solving this problem leads to the *solution*. Plausible, but incorrect, answers to the problem are referred to as *distractors*. The solution and distractors are used to create the *options*, or choices presented that the student must choose from.

Numerous guides for constructing multiple-choice questions exist and largely agree on the best practices for developing assessments (Moreno et al. 2015; Frey et al. 2005). These guides are routinely used by content specialists to create multiple-choice items, which are then disseminated for general use. Guidelines commonly focus on writing the content and choices of an item. For example, Haladyna and Rodriguez (2013) proposed 31 suggestions when writing multiple-choice items: 8 related to content and 14 related to choices. These suggestions are often times vague (e.g., "avoid trick or ambiguous items") and do not provide a way to

Darryl Chamberlain Jr.
University of Florida
E-mail: dchamberlain31@ufl.edu

Russell Jeter
Emory University
E-mail: rjeter@emory.edu

[Stem]
Solve the linear equation below.

[Problem]
$$\frac{-3x-6}{3} - \frac{-8x-8}{5} = \frac{7x+6}{2}$$

[Options]

A. $x = -\frac{40}{29}$ **[Distractor]** B. $x = -\frac{66}{29}$ **[Distractor]**

C. $x = -\frac{34}{29}$ **[Solution]** D. $x = -\frac{17}{10}$ **[Distractor]**

Figure 1: Example of a typical multiple-choice item.

systematically develop multiple-choice items. In fact, the authors state "The science of MC item writing is advancing, but item writing is still largely a creative act" (p. 329). The development of a systematic guide to create distractors based on common errors and misconceptions would provide an avenue to advance multiple-choice item writing in a "non-creative" way.

There are two common goals when utilizing multiple-choice assessments: summative and formative. *Summative* assessments strive to record student achievement as a capstone to some learning. Scores to these assessments are meant to provide a "summary" of each student's learning. In contrast, *formative* assessments strive to gather evidence of student learning in order to modify instruction. Scores to these assessments are meant to provide diagnostic information and inform the instructor what each student knows or does not know over some area of content (Cauley and McMillan 2010). A common criticism of multiple-choice assessments is that they cannot provide insight into student thinking (Lissitz et al. 2012) and thus multiple-choice assessments are routinely only utilized in a summative role. This is a function of two separate issues: (1) distractors must emulate common student conceptions or errors and (2) multiple-choice assessments cannot guarantee a student chose an option because it reflected their knowledge. In other words, a multiple-choice assessment that could provide diagnostic inferences about student thinking would need to be able to "predict" students conceptions with well-designed distractors AND ensure their option choice reflects their knowledge. We turn to technology to aid in this endeavor.

2 Automated Item Generation

One avenue for creating distractors based on conceptions and common misconceptions is Automatic Item Generation (AIG). AIG utilizes computer technologies and content specialists to automatically generate problems, solutions, and distractors. By automatically, we mean that an item structure template can be developed ahead of time that some technology would use to create many items without the need for future human intervention. Our method is an extension of the method introduced by Gierl and Lai (2013). An overview of our method is provided below, while a more detailed version of the method was presented in Chamberlain Jr. and Jeter (2020).

Step 1: Item Generation
In this step, the stem-type should be determined. This is equivalent to writing a free-response question and must include the stem and problem. In order to procedurally-generate versions of the question, elements of the stem and problem that can be modified must be identified at this point. A 1-layer model should be developed if only some small number of elements in the model can be modified. An n-layer model should be developed if many elements at multiple levels in a model can be modified (Gierl and Lai 2013). To be clear - this step requires careful **human** planning to create a free-response question that can be procedurally-generated. We are not suggesting a computer be able to develop and create the stem and mathematical problem of the question.

Step 2: Content
In this step, the content knowledge required to solve the problem is determined. To accommodate the de-

velopment of plausible distractors, any common errors or misconceptions associated to the problem should also be determined here. This can be collected by content specialists recalling common errors or misconceptions they are familiar with, recording any common errors identified in educational research experiments, or theoretically-predicted errors or misconceptions according to published mathematics education theoretical perspectives (Chamberlain Jr. and Jeter 2019).

Step 3: Computer-based Algorithms
In this step, the content knowledge collected in step 2 is utilized to procedurally solve the problem. In addition, distractor solutions can also be generated by:

1. Isolating common conceptual misunderstandings or common errors related to the topic assessed by the problem.
2. Using these misunderstandings and/or errors to construct "nearby problems".
3. Algorithmically solving these nearby problems to create a list of distractor solutions.

If the solution and distractor solutions are numeric in nature, the options can be disguised by algorithmically generating intervals that must satisfy two criteria:

1. There is minimal overlap between intervals (as any overlap will not contain a solution).
2. The intervals do not reveal much information about the solutions they are disguising.

To create distinct nearby problems based on common misconceptions or errors, the original stem/problem may need to be modified or a check may need to be created to regenerate the question until common misconceptions or errors do not produce the same solution as the correct solution (Chamberlain Jr. and Jeter 2020).

3 Implementation in College Algebra

The context of this study is a College Algebra course coordinated at a large, southeastern university in the United States. In Fall 2017, before the method was introduced, we used a question bank provided by Pearson to create multiple-choice exams. Items in these exams had 4 options: 3 distractors and the correct answer. In the initial implementation of the automated item generation method in Spring 2018, the 'content' step utilized theoretically-predicted errors and misconceptions based on our experience teaching the course. Each multiple-choice question was designed to have 5 options and disguised numeric options as necessary. After the initial implementation in Spring 2018, we designed a study to both (1) compare the new automated assessments to the previous assessments and (2) analyze the effects and validity of the automated assessments in and of themselves. We focused this exploratory analysis of the automated assessments around the following research questions:

Research Questions

1. Did the theoretically-driven distractors make for effective distractors?
2. How did the automated assessment generation change student performance distribution on similar exams?
3. How well or poorly did the automated items predict student performance on the exams?
4. How consistent were the automatically generated assessments?
5. How did the automatically generated assessments change the validity on similar exams?

Throughout the content changes the coordinator implemented in the course redesign that accompanied the integration of the automated assessment generation, the final exam content remained relatively the same. Therefore, we performed quantitative and qualitative analyses of each choice, item, and/or version of the Final Exam during Fall 2017, Fall 2018, and Fall 2019. The semester remained constant as different populations of students take the course during different semesters. The coordinator did not have item data for Final Exams taken during Fall 2016 or before, and thus Fall 2017 was chosen as the first semester to analyze the assessments. The number of students that tested on each version of the exam is provided in Table 1.

	Automated Item Generation	Ver A	Ver B	Ver C	Total
Fall 2017	None	76	75	74	225
Fall 2018	Theoretically-driven	69	70	65	204
Fall 2019	Theoretically- and Experimentally-driven	70	57	60	187

Table 1: Automated Item Generation method and number of students who took each version of the Final Exam.

4 Quantitative Comparison of Fall 2017, 2018, and 2019

Units of analysis and appropriate statistical measures were chosen to correspond to each research question individually. These were:

- Percent of Items with X many effective distractors;
- Percent of distractors chosen at $X\%$;
- Item difficulty distribution;
- Point Biserial Correlation; and
- KR-20.

Percent of Items with X many Effective Distractors (% X-ED) measured the number of items with 0, 1, 2, 3, or 4 effective distractors against the total number of items in each Final Exam. An effective distractor is one that is chosen by students at a rate of at least 5% (Hingorjo and Jaleel 2012). For example, 21% 1-ED means that 21% of items had exactly one effective distractor. Ideally, every distractor would be chosen at least 5% of the time for every question. 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. Given the focus of data in this paper on final exam data, we consider an item to be poor if there were 0 effective distractors, okay if there was 1 effective distractor, and good if there are 2+ effective distractors.

Percent of Distractors Chosen at X % (% D at X %) measured the rate at which students chose a distract against the total number of distractors on the exam. For example, 7% D at 0% means that 7% of distractors for the exam were not chosen by students. The categories 0%, 0%-5%, and 5%+ were used to illustrate poor distractors (those not chosen by students), distractors that were either choosen at random or elicited a small percentage of students (potentially effective distractors), and those that were effective. We consider the category of 0%-5% to be important as these distractors could potentially be refined or modified to elicit 5%+ student choice in a future exam. In other words, these would be the target distractors to possibly modify, while the category 0% would be target distractors to possibly eliminate.

Item Difficulty Distribution measured the distribution of the individual item difficulties (rates students answered the question correctly) as an entire test. This included categorizing the item distributions by 0.1 to consider the "shape" of the data (e.g., normal, skewed, uniform) as well as considering the mean, median, standard deviation, skewedness, and kurtosis of the item difficulties. Literature on assessments in general suggest item difficulty should be normally distributed with as small a standard deviation as possible (Lord 1952) while education assessment literature notes this is not commonly obtained in practice (Ho and Yu 2015). With both in mind, we analyzed the distribution in terms of the following ideals:

- Center: 0.7 (5-option) or 0.74 (4-option)
- Standard Deviation: 0.1 (5-option) or 0.087 (4-option)
- Skew and Kurtosis: 0

The first statistic, center, was directly suggested by Lord (1952). We considered both the mean and median when analyzing the center of the data, as non-normal distributions (such as those likely to be obtained

in educational assessment) can have a large difference between mean and median. The values proposed for the second statistic, standard deviation, were suggested as ideal in that 3 standard deviations from the ideal center would contain 99.7% of the ideally normalized data. The third and fourth statistics, skew and kurtosis, are suggested as ideal as 0 for each would indicate a normal distribution. Both statistics are used as skewness measures the asymmtery of the data (and thus describes the peak of the distribution) while kurtosis generally describes the tails of the distribution (Ho and Yu 2015). Taken together, the mean, median, standard deviation, skew, and kurtosis provide a detailed description of the shape of the distribution and thus can provide insight into how reliably the assessment describes student understanding.

Point Biserial Correlation (PBC) measures the the correlation between a student's answer on a particular question and their score on the assessment as a whole. Positive PBC corresponds to high-achieving students marking the question correctly while low-achieving students mark the same question incorrectly. The strength of this correlation ranges from -1 to 1, where PBCs 0 denote little correlation between student achievement on the exam as a whole and their success on the particular question and PBCs close to -1 or 1 denote a strong correlation. As such, items with a 0 or negative PBC are poor and should be revised. Items with a PBC between 0 and 0.15 have a weak correlation with overall score and should be revised. Items with a PBC score of 0.15 or over are considered acceptable, though PBCs of 0.25 or above are considered good (Varma 2006).

Finally, the KR-20 reliability coefficient is used to estimate the internal consistency reliability of an assessment (Salvucci et al. 1997). In other words, the KR-20 coefficient attempts to measure whether another group of similar students achieve in a similar way. The authors suggested the following criteria:

- Less than 0.5, the reliability is low;
- Between 0.5 and 0.8, the reliability is moderate; and
- Greater than 0.8, the reliability is high (p. 115).

This section will focus on research questions 1 through 4 individually, then summarize the results of each measure to provide a holistic consideration of the quantitative evaluation of the assessment.

4.1 Effective distractors

The first research question focused on whether or not the theoretically-drive distractors made for effective distractors. The literature suggests that *effective* distractors are those that are chosen at least 5% of the time (Hingorjo and Jaleel 2012). To answer this question, we analyzed both the items and the distractors as units by considering how many effective distractors an item had and what percentage of distractors were effective. Results for Fall 2017, 2018, and 2019 are presented in Tables 2 and 3.

	Fall 2017				**Fall 2018**				**Fall 2019**			
	Ver A	Ver B	Ver C	AVG	Ver A	Ver B	Ver C	AVG	Ver A	Ver B	Ver C	AVG
Items w/ 0 EDs	12%	20%	32%	21%	17%	17%	13%	15%	9%	5%	0%	5%
Items w/ 1 EDs	32%	24%	32%	29%	33%	29%	25%	29%	27%	41%	41%	36%
Items w/ 2 EDs	28%	32%	28%	29%	17%	33%	25%	25%	45%	32%	23%	33%
Items w/ 3 EDs	28%	24%	8%	20%	17%	8%	25%	17%	18%	14%	27%	20%
Items w/ 4 EDs	NA	NA	NA	NA	8%	4%	4%	6%	0%	9%	9%	6%

Table 2: Percentage of items with $0, 1, 2, 3,$ or 4 effective distractors (EDs) in Final Exams during Falls 2017, 2018, and 2019.

Note the progression in percentage of items with 0 effective distractors: 21%, 15%, 5%. This is clear evidence that through the design and implementation of the automated item generation, the quality of distractors *per item* significantly increased. Moreover, as the AIG method developed 4 distractors for each item, a small percentage of items (6% average) produced all quality distractors.

Again note that Fall 2017 had 3 distractors per question while Fall 2018/2019 had 4 distractors per question. Even with the increase in distractors, distractor choice percentages stayed relatively similar between

	Fall 2017				Fall 2018				Fall 2019			
	Ver A	Ver B	Ver C	AVG	Ver A	Ver B	Ver C	AVG	Ver A	Ver B	Ver C	AVG
Ds chosen 0%	7%	17%	21%	15%	21%	22%	19%	20%	19%	27%	15%	20%
Ds chosen 0%–5%	36%	29%	41%	36%	38%	43%	38%	39%	38%	27%	34%	33%
Ds chosen 5%+	56%	53%	37%	49%	42%	34%	41%	39%	43%	45%	50%	46%

Table 3: Percentage of distractors (Ds) that were chosen 0%, 0%–5%, or 5%+ of the time by students in Final Exams during Falls 2017, 2018, and 2019.

Fall 2017 and Fall 2019. This again suggests the automated assessment generation created more effective distractors that were well-distributed between items.

Overall, the data suggests the theoretically-driven distractors made for effective distractors. The percentage of items with at least one effective distractor progressed from 79%, to 85%, and finally to 95%. Moreover, the increase in the number of distractors generated (from 3 to 4) did not substantially change the percentage of distractors chosen by students.

4.2 Student performance on exam

Student performance on the exams was analyzed by the *item difficulty*, or the proportion of students answering each item correctly. This analysis consisted of considering 5 key statistics: (1) mean, (2) median, (3) standard deviation, (4) skew, and (5) kurtosis. Theoretically, multiple-choice assessments are most predictive when results are perfectly normally distributed around 0.7 (5-options) or 0.74 (4-options) (Lord 1952). When average item difficulty is low, all students with adequate knowledge are separated from the lowest achieving students, but the assessment would have trouble separating achievement within the larger group of successful students. When average item difficulty is high, it is more likely students will randomly guess rather than answer according to their own understanding. This is why assessments attempt to strike a balance of questions with item difficulties between 0.5 and 0.8.

We did not perform a Kolmogorov-Smirnov test or Shapiro-Wilk test to determine whether the data was normally distributed as educational assessments do not regularly attain the idealized normal distribution of assessments in other fields such as psychiatric or medicine Ho and Yu (2015). Instead, we considered the statistics to holistically describe the distributions as degrees from normality. All statistics were calculated using built-in Excel functions. We assume the reader is familiar with mean, median, and standard deviation and thus will provide a short description of skewness and kurtosis.

Skewness is a measure of the departure from a horizontal symmetry of a peak Bulmer (1979). The sign of the skewness suggests the direction of the skew: negative as left-skewed and positive as right-skewed. The magnitude of the skewness suggests how skewed the distribution is, according to the following categories:

- *Highly skewed* - Magnitude of the skewness is greater than 1;
- *Moderately skewed* - Magnitude of the skewness is between 1 and 0.5; and
- *Approximately symmetric* - Magnitude between 0.5 and 0 Bulmer (1979).

Kurtosis is a measure of the tails of a distribution (Ho and Yu 2015). Like Excel, the kurtosis values we present are actually the *excess kurtosis*, or difference of kurtosis as compared to the kurtosis of a normal distribution (3). As such, a negative kurtosis suggests more data to appear in the tails of the distribution (a uniform distribution has a kurtosis of -1.2) and a positive kurtosis suggests little data to appear in the tails (a logistic distribution has a kurtosis of 1.2).

Together, skewness and kurtosis can provide information about the simiarlity of a curve to a normal distribution. Skewness considers the symmetry of the distribution while kurtosis considers the general shape of the tail to the peak. For example, a skewness near 0 *and* kurtosis near 0 suggests a normal distribution, while a skewness near -1 (or 1) and kurtosis near 1 suggests a skewed distribution. However, a skewness near -1 or 1 and kurtosis near 0 suggests the data is skewed due to a lower or upper bound for the data, such as students not being able to score lower than a 0 or higher than a 100 on an exam.

In summary, we analyzed the item difficulty distributions against the following “practical” ideals:

- Center: 0.7 (5-option) or 0.74 (4-option)

– Standard Deviation: 0.100 (5-option) or 0.087 (4-option)
– Skew and Kurtosis: 0

While the ideal center was suggested by Lord (1952), the other statistics are meant to provide a frame of reference when discussing the item distribution beyond the overall look of the distribution provided in Figures 2 and 3. This allows us to determine whether the assessments, in terms of the item difficulty, are improving their reliability by moving toward these practical ideals from their values in Fall 2017.

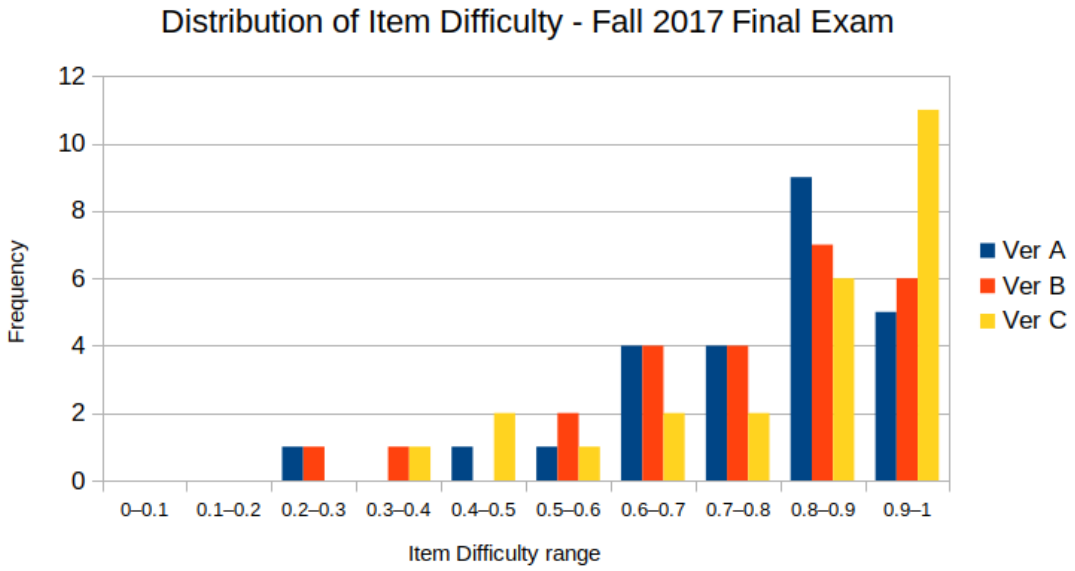

Figure 2: Item Difficulty Distribution for the Final Exam in Fall 2017.

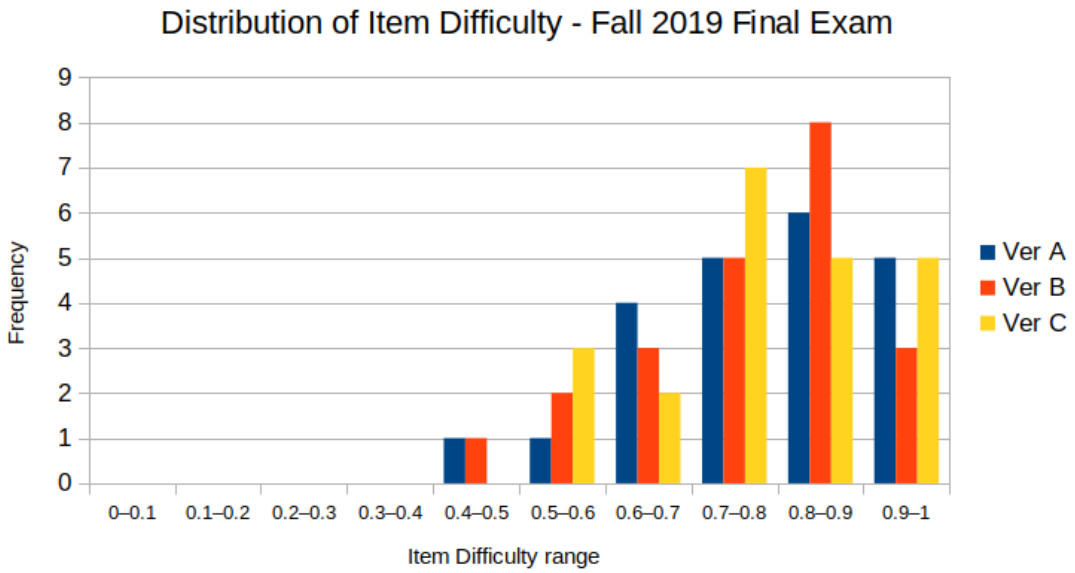

Figure 3: Item Difficulty Distribution for the Final Exam in Fall 2019.

A cursory look at Figures 2 and 3 suggests the item difficulty distributions moved from a highly skewed-left distribution to a moderately skewed-left distribution. A summary of their item difficulty statistics is provided in Table 4 that confirms the general distribution shapes and provides a more nuanced comparison

between the versions. We will focus on comparing the distributions of Fall 2017 and Fall 2019 but included summary statistics for Fall 2018 for completeness.

Table 4: Summary of item difficulty distribution statistics for Fall 2017, 2018, and 2019.

	Fall 2017				**Fall 2018**				**Fall 2019**			
	Ver A	Ver B	Ver C	AVG	Ver A	Ver B	Ver C	AVG	Ver A	Ver B	Ver C	AVG
Mean	0.76	0.77	0.81	0.78	0.76	0.83	0.74	0.77	0.77	0.77	0.76	0.77
Median	0.82	0.80	0.89	0.83	0.78	0.83	0.75	0.78	0.79	0.80	0.78	0.79
Std Dev	0.168	0.181	0.170	0.174	0.150	0.149	0.158	0.155	0.121	0.136	0.122	0.127
Skewness	-1.64	-1.15	-1.20	-1.33	-0.39	-1.63	-0.56	-0.86	-0.71	-0.96	-0.71	-0.79
Kurtosis	3.46	0.97	0.62	1.68	-1.15	4.05	0.43	1.11	0.05	0.46	-0.47	0.02

All 5 statistics for Fall 2017 suggest the item difficulty distribution was highly skewed-left. The median score was on average 0.05 higher than the average mean, suggesting a majority of data exists to the right of the mean. The average skewness was -1.33, which is above the threshold of -1 for highly skewed distributions. Each individual kurtosis was above 0.5, which suggests smaller tails with a larger peak that are common among skewed distributions. Beyond establishing the distribution as highly skewed-left, the center of 0.83[1] is significantly above the ideal of 0.74 and the average standard deviation of 0.174 is significantly above the ideal of 0.087. The distribution suggests the assessment is not as reliable as it could be in distinguishing between students' achievement levels.

In contrast to Fall 2017, the statistics for Fall 2019 suggest the item difficulty distribution may have been moderately skewed-left due to the cap in score a student could achieve. Note the average skewness of -0.79 and average kurtosis of 0.02 – this suggests the left side of the distribution had a similar tail/peak as a normal distribution but the right side of the distribution was cut off by the inability to score above a 100% on the exam. Beyond establishing that the distribution is at least moderately skewed-left, the center and spread of the data were closer to the practical ideals. The center of the distribution, 0.77, was not close to the ideal center of 0.70, but the average standard deviation of 0.127 was somewhat close to the ideal of 0.100. The distribution suggests the assessment is approaching the quantitative statistics that would suggest high reliability in distinguishing between students' achievement levels.

Overall, we see a reduction in the skewedness and spread of the item difficulty distribution from Fall 2017 to Fall 2019. This suggests that the automated assessment generation was superior at generating an exam that could distinguish between varying student levels of achievement rather than just the most extreme.

4.3 Predict student performance

The previous section spoke to the reliability of an assessment in terms of the overall assessment scores of students. We can also consider how predictive a single item is of a student's overall score. The Point Biserial Correlation (PBC) measures exactly this. Positive PBC corresponds to high-achieving students marking the question correctly while low-achieving students mark the same question incorrectly, while negative PBC corresponds to the inverse relation. Therefore, items with high PBC are strong predictors of student achievement while low PBC indicates a potentially poorly-written item. PBCs were categorized according to Varma (2006) and can be found in Table 5.

Unlike the previous sections, all three exams performed relatively the same when it comes to predictive questions. We also see strong variation between versions, even those that were not automatically generated. Considering the visual outliers, we see that Version B in Fall 2018 had 25% of questions in the PBC range of 0-0.15. Looking back at Table 4, we see this version also had a far higher than average of 0.83 and the largest kurtosis of 4.05. Together, this highlights the danger of items with low item difficulty mentioned in the item distribution section – these items may result in less predictability in overall scores. These issues seemed to be addressed in the next iteration of the Final Exam, and thus it appears that both assessment types were able to provide similar rates of highly predictive items.

[1] Median is a better estimate of center in a skewed distribution.

	Fall 2017				Fall 2018				Fall 2019			
	Ver A	Ver B	Ver C	AVG	Ver A	Ver B	Ver C	AVG	Ver A	Ver B	Ver C	AVG
PBCs $\leq$ 0	0%	0%	4%	1%	0%	4%	0%	1%	0%	0%	0%	0%
PBCs 0–0.15	8%	8%	12%	9%	0%	25%	4%	10%	5%	14%	0%	6%
PBCs 0.15–0.25	4%	13%	16%	11%	17%	8%	17%	14%	0%	14%	18%	11%
PBCs $\geq$ 0.25	88%	83%	72%	81%	83%	63%	79%	75%	95%	73%	82%	83%

Table 5: Percentage of items with a Point Biserial Correlation (PBC) in Final Exams during Falls 2017, 2018, and 2019.

4.4 Assessment consistency

Finally, the Kuder–Richardson Formula 20 (KR-20) reliability coefficient is used to estimate the internal reliability of an assessment (Salvucci et al. 1997). A high[2] KR-20 suggests the assessment would highly correlate between alternative forms, something that is especially desirable when alternative forms are automatically generated. We provide the KR-20 for all versions of the Final Exams in Fall 2017, 2018, and 2019 in Table 6.

	Fall 2017				Fall 2018				Fall 2019			
	Ver A	Ver B	Ver C	AVG	Ver A	Ver B	Ver C	AVG	Ver A	Ver B	Ver C	AVG
KR-20	0.799	0.792	0.599	0.730	0.736	0.614	0.735	0.695	0.727	0.654	0.710	0.697

Table 6: KR-20 for Final Exams in Fall 2017, 2018, and 2019.

Similar to the section on PBC, we see relatively the same KR-20s through the semesters with some variation within versions. As all KR-20s are within the moderately reliable range for an assessment, it appears that the automatically generated assessments are just as internally consistent as the Fall 2017 assessments.

4.5 Summary of Quantitative Results

For ease of comparison across multiple statistics, we have summarized the 5 statistics, their definitions, how we rated them, and their average for Final Exams in Fall 2017, 2018, and 2019 in Table 7.

Statistically, the automatically generated Final Exam of Fall 2019 outperformed the item bank Final Exam of Fall 2017 in two particular measures: percent of items with effective distractors and item difficulty distribution. Fall 2019 showed a significant decrease in the percentage of items with no effective distractors (21% to 5%) as well as the corresponding increases in items with 1 (29% to 36%) or 2+ (49% to 59%) effective distractors. This is strong evidence that the theoretically-drive distracotrs of the automated assessment generation produced effective distractors. Moreover, the Final Exam in Fall 2019 produced a truncated normal distribution with smaller standard deviation (0.127 versus 0.174) than Fall 2017. This suggests the automated assessment generation produced more predictive assessments of student success as they aligned realistically close to best assessment practices for item difficulty distribution.

5 Content and Option Validity

The previous section focused on a statistical analysis of the assessments, considering statistically effective distractors, item difficulty distribution, and correlation coefficients meant to evaluate item and assessment *reliability*, or the consistency of the items and assessment. This section will switch to a qualitative approach of evaluating the *validity* of the assessments, or whether the items and options measured what they proported to measure.

[2] It is important to note that too high of a KR-20, usually above 0.9, is actually not desirable in educational assessment as it could indicate redundancy of items or items with no ability to discriminate between individuals.

			2017	2018	2019
Measure	**Rating**	**Description**	**AVG**	**AVG**	**AVG**
Effective Distractors *Distractors chosen at 5%+.*	Poor	Items w/ 0 EDs	21%	15%	5%
	Okay	Items w/ 1 EDs	29%	29%	36%
	Good	Items w/ 2+ EDs	49%	47%	59%
Distractor Chosen *Overall percentage chosen.*	Poor	Distractors chosen 0%	15%	20%	20%
	Okay	Distractors chosen 0%–5%	36%	39%	33%
	Good	Distractors chosen 5%+	49%	39%	46%
Item Difficulty *Proportion of students answering each item correctly.* *Mean and Median describe the center of the distribution. Ideal: 0.7 or 0.74.* *Standard Deviation describes the spread of the distribution. Ideal: 0.100 or 0.087.* *Skew describes the symmetry of distribution. Ideal: 0.* *Kurtosis describes the shape of the tail to the peak of a distribution. Ideal: 0.*	Poor	Statistics suggest non-normal distribution. **Mean** **Med** **StDev** **Skew** **Kurt**	0.78 0.83 0.174 -1.33 1.68		
	Okay	Statistics suggest somewhat normal distribution near "good" center and standard deviation. **Mean** **Med** **StDev** **Skew** **Kurt**		0.77 0.78 0.155 -0.86 1.11	
	Good	Statistics suggest normal distribution near "good" center and low standard deviation. **Mean** **Med** **StDev** **Skew** **Kurt**			0.77 0.79 0.127 -0.79 0.02
Point Biserial Correlation *Correlation between score on item and total score.*	Poorest	Below 0	1%	1%	0%
	Poor	Between 0 and 0.15	9%	10%	6%
	Okay	Between 0.15 and 0.25	11%	14%	11%
	Good	0.25 and above	81%	75%	83%
KR-20 *Correlation of alternative forms.*	Poor	0.5 or below; Between 0.9 and 1.0			
	Good	Between 0.5 and 0.8	0.730	0.695	0.697
	Ideal	0.80 - 0.9			

Table 7: Summary of averaged quantitative results for Final Exams during Fall 2017, 2018, and 2019.

Content validity refers to an item measuring the content it is intended to measure. Within the educational sphere, this normally refers to whether an item measures the content objectives attached to the question at hand. Questions in an exam bank from online platforms normally are tagged with one or more content objectives aligned to the textbook. Content validity is normally the first concern when developing questions for an assessment (Allen and Yen 2001).

The course coordinator reviewed the content objectives attached to each question of the Final Exams administered in Fall 2017, 2018, and 2019. Questions with the same content objectives were compared to identify if the content validity was improved by the automated item generation. For the majority of "solve the equation" type questions, content validity improved from Fall 2017 to Fall 2018/2019. We provide exemplary proof in Table 8.

On the surface, each question in Table 8 fits the description of the content objective "Solve linear functions with rational coefficients" – they are all linear equations with rational coefficients after all. However, there are two issues with the Fall 2017 question that correlate with a student not **solving** the linear equation:

(a) A student could simply check whether each fraction makes the equality true. If a student does this, the associated content objective becomes "reduce compound fractions to show equivalence", which is distinct from the tagged content objective.
(b) A student could take an educated guess that the correct answer, D, is the one that shares either a numerator or denominator with each of the other options. If a student does this, there is no associated content objective.

These alternative explanations for why a student may choose the correct answer bring the content validity of the Fall 2017 item into question. Sidestepping the mathematical question may also explain why the percentage of correct responses is far higher (91%) than the ideal rate (74%) and no distractor is effective. This suggests we have evidence the Fall 2017 question is not measuring students' ability to solve a linear equation with rational coefficients.

Table 8: Final Exam question from Fall 2017-2019 that tests on the course objective "Solve linear equations with rational coefficients."

	Fall 2017 Q9	**Fall 2018 Q9**	**Fall 2019 Q5**
Problem	$\frac{x-2}{9}=\frac{x-4}{2}$	$\frac{-4x-6}{2}-\frac{-4x+6}{5}=\frac{3x+7}{4}$	$\frac{-8x+7}{4}-\frac{-3x+4}{5}=\frac{-3x+6}{2}$
Options	A. $\left\{\frac{40}{7}\right\}$ B. $\left\{-\frac{32}{11}\right\}$ C. $\left\{\frac{34}{7}\right\}$ *D. $\left\{\frac{32}{7}\right\}$	A. $x \in [-1.9, -0.7]$ *B. $x \in [-5.1, -2.1]$ C. $x \in [-11.6, -8.8]$ D. $x \in [-3, -1.9]$ E. There are no Real solutions.	A. $x \in [2, 5]$ B. $x \in [27, 31]$ C. $x \in [-5, 0]$ *D. $x \in [18, 25]$ E. There are no Real solutions.
Percentage of Responses	4%, 3%, 3%, **91%**	20%, **68%**, 3%, 4%, 4%	10%, 3%, 17%, **66%**, 4%

Students could potentially sidestep the content objective for Fall 2018/2019 by solving the equation equal to 0, plugging in the endpoints of each interval, and utilizing Intermediate Value Theorem to determine which interval holds the solution. Not only is this process more time-consuming than the two issues described for the Fall 2017 item, but it also requires knowledge typically acquired *far beyond a College Algebra course*. It is therefore unlikely a student would sidestep the mathematical question in this manner. Moreover, consider the percentage of responses for the same structure in Fall 2018 and 2019. Both questions had a correct response rate around 67%, suggesting high reliability between similar sets of students. Option A in both questions also corresponds to the same distractor: improperly distributing the subtraction sign in front of the second fraction. While the rates are not as close as those of the correct responses, they are both above the 5% threshold that suggests the distractor is effective. As the correct response rate is close to the ideal rate of 70%, we have strong evidence that the questions in Fall 2018/2019 are measuring students' ability to solve a linear equation with rational coefficients.

In Table 9, all rates of correct responses are far above the ideal rate: 91% versus 74% for Fall 2017 and 87%/83% versus 70% for Fall 2018/2019. When we control for the Fall 2017 problem not additionally testing on strict inequality, we see the correct responses rates become even closer: 91% versus 88%/87%. Similarly in Table 10, Fall 2018 students correctly responded to the factoring the quadratic *at a higher rate* than Fall 2017: 93% versus 86%. These two problems suggest that even though students can sidestep the mathematical problem, they may perform the same quantitatively on a multiple-choice item. Combined with the previous example, these questions illustrate that multiple-choice options can undermine the content valdity of a mathematical problem.

The previous examples illustrate questions that would have content validity *if they were a free-response question*, but do not as-written. The automated item generation method, and specifically the interval generation method, allow the same types of problems to preserve their content validity as multiple-choice questions. Within content validity, we can consider *option validity*, which refers to whether the option response rate measures what it purports to measure. This could provide evidence of students who circumvent the mathematical problem by considering what a student would need to do in a problem to choose a distractor. Consider the following example below.

One of the most common mistakes a student could make solving linear equations with integer coefficients is not distributing the negative. If a student consistently does this for the question in Table 11, their answer would be -49. Yet more students chose $-\frac{49}{2}$, the similar manipulation of this misconception, than students who chose the misconception itself! It is not clear how a student could work through the problem and get $\frac{-49}{2}$ (or $\frac{21}{2}$ for that matter) and thus we have evidence the question lacks option validity. This suggests that

Table 9: Final Exam question from Fall 2017-2019 that tests on the course objective "Solve a compound linear inequality."

	Fall 2017 Q12	**Fall 2018 Q12**	**Fall 2019 Q4**
Problem	$-1 \leq \frac{x+1}{3} \leq 3$	$7+7x < \frac{50x-4}{6} \leq 8+8x$	$-8+8x < \frac{36x-8}{4} \leq 7+8x$
Options	A. $[-5,3]$	A. $(a,b]$, where $a \in [-29,-25]$ and $b \in [-10,-5]$	A. $[a,b)$, where $a \in [-12,-7]$ and $b \in [3,7]$
	B. $[-7,1]$	B. $[a,b)$, where $a \in [-29,-24]$ and $b \in [-10,2]$	B. $(a,b]$, where $a \in [-10.4,-8.6]$ and $b \in [4.6,6.4]$
	C. $[-1,7]$	*C. $(a,b]$, where $a \in [2,8]$ and $b \in [23,31]$	*C. $(a,b]$, where $a \in [-8.2,-4.9]$ and $b \in [6.1,11.2]$
	*D. $[-3,5]$	D. $[a,b)$, where $a \in [4,7]$ and $b \in [25,29]$	D. $[a,b)$, where $a \in [-7,-2]$ and $b \in [8,11]$
		E. There is no solution to the inequality.	E. There is no solution to the inequality.
Percentage of Responses	7%, 1%, 1%, **91%**	1%, 3%, **87%**, 3%, 6%	0%, 4%, **83%**, 10%, 3%

Table 10: Final Exam question from Fall 2017-2019 that tests on the course objective "Factor a quadratic function with $a > 1$."

	Fall 2017 Q7	**Fall 2018 Q6**
Problem	Factor $-24x^2 - 20x + 24$	Factor $15x^2 + 62x + 40$ as $(ax+b)(cx+d)$, with $b \leq d$.
Options	A. $-4(3x+2)(2x-3)$	A. $a \in [-4.1,-2.6], b \in [-15,-3], c \in [-7,-3]$, and $d \in [-8,-2]$
	B. $4(3x+2)(2x-3)$	B. $a \in [0.7,1.6], b \in [-3,7], c \in [13,17]$, and $d \in [8,13]$
	*C. $-4(3x-2)(2x+3)$	*C. $a \in [4.3,7.8], b \in [-3,7], c \in [-3,6]$, and $d \in [8,13]$
	D. $4(x+1)(x-2)$	D. $a \in [0.7,1.6], b \in [9,13], c \in [13,17]$, and $d \in [-1,9]$
		E. $a \in [-4.1,-2.6], b \in [9,13], c \in [-7,-3]$, and $d \in [-1,9]$
Percentage of Responses	13%, 1%, **86%**, 0%	0%, 1%, **93%**, 6%, 0%

even if distractors are made with common student misconceptions and/or errors, students may not choose these options if there are ways to sidestep the mathematical problem.

Unlike the previous example, our method to generate items provides strong evidence students' choices are associated to their understanding and not the appearance of the options. Consider question five from the final exam in Fall 2019 associated to the objective "Solve linear equations with rational coefficients" in Table 12.

Table 11: Final Exam question from Fall 2017 that tests on the course objective "Solve a linear equation with integer coefficients."

	Fall 2017 Q8
Problem	$-7[2x+5-3(x+1)]=6x+7$
Options	A. -49 B. $\frac{21}{2}$ C. $-\frac{49}{2}$ *D. 21
Percentage of Responses	0%, 4%, 5%, **91%**

Table 12: Final Exam question from Fall 2019 that tests on the course objective "Solve a linear equation with rational coefficients."

	Fall 2019 Q5
Problem	$\frac{-3x+3}{5}-\frac{-3x-4}{6}=\frac{-3x-6}{2}$
Options	*A. $x \in [-3.76, -2.29]$ B. $x \in [-2.24, -1.87]$ C. $x \in [-2.04, -0.58]$ D. $x \in [-9.91, -9.07]$ E. There are no real solutions
Explanation of Distractors	A. Correct option. B. Not distributing the negative in front of the second fraction to both terms in the numerator. C. Dividing only the second coefficient in each numerator by the denominator. D. Dividing only the first coefficient in each numerator by the denominator. E. Believing it was not possible to solve the equation.
Percentage of Responses	**77%**, 14%, 7%, 0%, 2%

The most common error, not distributing the negative in front of the fraction, was expected. What was not expected was that students would choose option C at the next highest rate and did not choose option D at all. In this particular type of question, students appeared to notice that $-6/2$ reduced in the third fraction, then applied this rule uniformly to the rest of the question. Once we noticed this, we modified the question structure generation to ensure that only one set of coefficients was divisible by at least one of the denominators to determine if students would be consistent in their choices on this question structure. As we expected, the pattern consisted during Spring 2020, as shown in Table 13. This is additional evidence that the options created by our method were associated to student understanding and not the display of the options.

Table 13: Exam 1 question from Spring 2020 that tests on the course objective "Solve a linear equation with rational coefficients."

	Spring 2020 Q10
Problem	$\frac{4x-3}{3} - \frac{-3x+6}{5} = \frac{3x+4}{2}$
Options	A. $x \in [2, 5]$ B. $x \in [-4, 2]$ *C. $x \in [8, 11]$ D. $x \in [28, 31]$ E. There are no real solutions
Explanation of Distractors	A. Not distributing the negative in front of the second fraction to both terms in the numerator. B. Dividing only the second coefficient in each numerator by the denominator. C. Correct option. D. Dividing only the first coefficient in each numerator by the denominator. E. Believing it was not possible to solve the equation.
Percentage of Responses	22%, 13%, **59%**, 0%, 6%

6 Discussion

To begin, we summarize the main takeaways from the quantitative and qualitative analysis:

- The theoretically-driven distractors utilized by AIG produced a plethora of effective distractors, resulting in an overall decrease from 21% to 5% of items with no effective distractors.
- AIG produced assessments with truncated normal item difficulty distributions and reduced standard deviations.
- Multiple-choice options can undermine the content validity of a mathematical problem by allowing students to circumvent the mathematical problem at hand.
- However, the interval mask method we employed through AIG allowed us to protect the content validity and option validity of multiple-choice items.

According to all analytical measures we utilized, the AIG method we utilized produced valid and reliable assessments. We believe this is due primarily to the theoretically-driven distractors and interval mask method. And yet the AIG also outperformed the item bank assessments in numerous practical considerations, such as:

- *Time efficiency*: It currently takes approximately 5 minutes to create 3 separate versions for questions attached to 80 different content objectives.
- *Formative efficiency*: By printing how each distractor was generated and how a student might overcome their error/misconception on a separate PDF, the automatically generated keys can be utilized as a powerful, individualized learning tool.
- *Cost/Portability efficiency*: The current scripts and code all utilize open-source tools such as Python and SageMath. This allows the assessments to be customized and utilized with any course materials at no cost to the instructor or students.

Speaking to the portability of the designed AIG content for College Algebra, we also found evidence that different populations of students may choose distractors at varying rates(as one might expect if the

options are associated to student understanding). Consider the percentage of items with effective distractors and total percentage of distractors chosen for the first exam in Fall 2019 between in-person students and online-only students in Table 14.

	In-Person				Online			
	Ver A	Ver B	Ver C	AVG	Ver A	Ver B	Ver C	AVG
Items w/ 0 EDs	10%	10%	21%	14%	5%	0%	0%	2%
Items w/ 1 EDs	40%	35%	37%	37%	25%	10%	21%	19%
Items w/ 2 EDs	30%	25%	26%	27%	10%	20%	16%	15%
Items w/ 3 EDs	10%	20%	11%	14%	25%	40%	16%	27%
Items w/ 4 EDs	10%	10%	5%	8%	35%	30%	47%	37%
Ds chosen 0%	10%	14%	16%	13%	6%	6%	7%	6%
Ds chosen 0%–5%	48%	40%	49%	45%	29%	21%	21%	24%
Ds chosen 5%+	43%	46%	36%	41%	65%	73%	72%	70%

Table 14: Summary of distractor percentages during Exam 1 in Fall 2019.

All students were given the exact same 3 versions. Even so, distractors were chosen at far higher rates by Online students. There is also a massive difference in percentage of questions considered effective for each group of students: 41% for Hybrid students and 70% for Online students. This suggests the quantitative measure of "chosen by at least 5% of students" may not provide the entire context for whether a distractor is truly effective, as this measure is heavily influenced by the understanding of students. This illustrates the need for theoretically-driven distractors *at the level of the students in the course and potentially between disparate groups of students within the same course* and challenges the recommendation of three to four options for maximum test efficiency (Cizek and O'Day 1994; Haladyna and Rodriguez 2013).

6.1 Utilizing Diagnostic Assessments

Recall that there are two common goals when utilizing multiple-choice assessments: summative and formative. Our method shows promise in countering the common criticism of multiple-choice assessments – that they cannot provide insight into student thinking (Lissitz et al. 2012). More than just provide insight into student thinking, our method provides a blueprint for how to modify instruction based on the students' answer choice through the use of targeted feedback.

A key component to the automated assessment generation is the development of theoretical distractors – distractors that align with common student misconceptions or errors. By associating a student's response with these distractors, we can provide targeted feedback that may help students overcome these misconceptions or errors. In fact, this process can be provided within the context of normal assessments in the course such as homework.

For our first example, consider the procedural content objective "Solve quadratic equations using the Quadratic Formula" in Figure 4. By including the common errors when writing the Quadratic Formula, a student would be able to quickly determine what they did incorrectly and thus learn from their mistake. Moreover, a common misconception that appeared through informal discussions with students after exams was the belief that if a quadratic equation could not be factored, it could not be solved. Both types of feedback are the same that would be given to students had they worked on the free-response version of the problem during an exam and so can effectively provide the targeted feedback needed for development.

Alternatively, we can control the structure of a question to potentially elicit common misconceptions and correct them. For example, consider two variations on the question structure associated to the objective "Solve a linear equation with rational coefficients."

$$\frac{-3x-3}{5}-\frac{-5x+9}{3}=\frac{5x+9}{7} \text{ and } \frac{-3x-9}{7}-\frac{8x-9}{2}=\frac{-9x-3}{4}$$

As illustrated in the qualitative section, we found evidence that the divisibility of the coefficients and denominators influenced students' solutions. Including both types of questions associated to the same structure

16. Solve the quadratic equation below. Then, choose the intervals that the solutions belong to, with $x_1 \leq x_2$ (if they exist).

$$-14x^2 + 9x + 9 = 0$$

The solution is $x_1 = -0.542$ and $x_2 = 1.185$

A. $x_1 \in [-19.3, -16.3]$ and $x_2 \in [7.56, 7.95]$

$x_1 = -16.593$ and $x_2 = 7.593$, which corresponds to using the Quadratic Formula with $a = 1$

B. $x_1 \in [-24.2, -23.8]$ and $x_2 \in [24.14, 24.86]$

$x_1 = -23.865$ and $x_2 = 24.508$, which corresponds to writing the Quadratic Formula as $-\frac{b}{2a} \pm \sqrt{b^2 - 4ac}$.

C. $x_1 \in [-1.1, 0.4]$ and $x_2 \in [1.08, 1.47]$

* $x_1 = -0.542$ and $x_2 = 1.185$, which is the correct option.

D. $x_1 \in [-1.3, -0.8]$ and $x_2 \in [0.19, 0.56]$

$x_1 = -1.185$ and $x_2 = 0.542$, which corresponds to writing the Quadratic Formula as $\frac{b \pm \sqrt{b^2 - 4ac}}{2a}$

E. There are no Real solutions.

Corresponds to getting a negative under the radical or believing that since the quadratic cannot be factored, it has no Real solutions.

Figure 4: Question for the content objective "Solve quadratic equations using the Quadratic Formula."

on an assessment, such as on homework, would allow the student to potentially confront and address this hidden misconception.

It is important to note that the association of responses to common errors and misconceptions can even improve online, free-response homework feedback. Rather than a wrong answer prompt asking the student to try again, the homework could prompt the student to directly address the associated error/misconception. The system could also keep track of recognized errors and misconceptions across homework problems that could trigger additional resources (e.g., videos or specialized walkthroughs) if a student continues to make the error or misconception. Incorporating these types of associations could be the next step in providing more dynamic just-in-time learning and effective feedback that is critical to student achievement in higher education (Robinson et al. 2015; Hattie and Timperley 2007).

7 Conclusions

Multiple-choice assessments are widely used in K-16 as summative tools. Utilizing technology, we propose a method to develop multiple-choice assessments that can be used as summative and/or formative tools. The quantitative analysis of exams generated utilizing our method illustrate that theoretically-driven distractors utilized by AIG produced a plethora of *effective distractors*, or distractors chosen by students at least 5% of the time. Moreover, the exam item difficulty distributions are normally distributed[3] with lowered standard deviation, which aligns with best practices for improved predictability of student success (Lord 1952). The qualitative analysis provided evidence that the use of intervals to mask numeric responses protected the content and option validity of the exam, allowing for a more reliable relation between a student's choice and their understanding of the content. The overall analysis showed that multiple-choice assessments generated using our method can provide insight into student thinking, countering the common criticism of multiple-choice assessments (Lissitz et al. 2012).

[3] The data suggests the distribution was actually a truncated normal distribution due to the maximum attainable score.

By aligning options to common misconceptions and errors in student thinking and providing strong evidence that students' choices align with their understanding, our method can be utilized to provide targeted, individualized feedback at little time-cost to the instructor that is pivotal to student achievement (Robinson et al. 2015; Hattie and Timperley 2007). Since item structures are created individually and then technology is utilized to generate hundreds of similar examples, these assessments can take the form of homework (giving students as much practice as they need with a concept) as well as exams (allowing instructors to make as many similar versions as necessary to protect exam security). This would potentially free up an instructor's time to administer more open-response assessments that cannot easily be automated. We see our automated item generation method as providing one avenue for improving K-16 education.

8 Limitations and Future Research

There are limitations to the automated item generation method we proposed and to the data analyzed. For one, the procedure is math-specific and likely not generalizable to Social Sciences. However, the procedure can be utilized in the sciences that require students to solve a mathematical problem. Being able to solve a mathematical problem is a fundamental procedure that spans multiple disciplines.

One semester of pre-automated assessment data was analyzed compared to two semesters of post-automated assessment data. Moreover, only the final exam from each semester was analyzed. While results would have been more convincing had more semesters and assessments been used, the presentation of this data was unwieldy and we decided to focus on a single assessment for three consecutive Fall semesters. This allowed for a comparison of similar students during the highest-enrollment semester to provide the most robust results. A future direction for research would be to compare student assessments throughout the semester to determine the effectiveness of automated assessments as diagnostic tools.

The items we presented were mostly of the "Solve the equation" type and are not conceptual. While Chamberlain Jr. and Jeter (2019) present how one may automatically generate conceptual questions, future work would include producing more automatically-generated conceptual questions.

Finally, there was a lack of qualitative interviews to confirm content validity. While we provide evidence to suggest student responses were associated with common errors and misconceptions, we do not provide evidence that their options matched their thinking. This is another avenue for future research.

Conflict of interest

The authors declare that they have no conflict of interest.

References

Allen MJ, Yen WM (2001) Introduction to measurement theory. Waveland Press

Bulmer M (1979) Principles of Statistics. Dover

Cauley K, McMillan J (2010) Formative assessment techniques to support student motivation and achievement. The Clearing House: A Journal of Educational Strategies, Issues and Ideas 81(1):1–6, DOI 10.1080/00098650903267784

Chamberlain Jr D, Jeter R (2019) Leveraging cognitive theory to create large-scale learning tools. In: Proceedings of the 22nd Annual Conference on Research in Undergraduate Mathematics Education, pp 741–747

Chamberlain Jr D, Jeter R (2020) Creating diagnostic assessments: Automated distractor generation with integrity. Journal of Assessment in Higher Education

Cizek G, O'Day D (1994) Further investigation of nonfunctioning options in multiple-choice test items. Educational and Psychological Measurement 54(4):861–872, DOI 10.1177/0013164494054004002

Frey B, Petersen S, Edwards L, Pedrotti J, Peyton V (2005) Item-writing rules: Collective wisdom. Teaching and Teacher Education 21:357–364, DOI 10.1016/j.tate.2005.01.008

Gierl M, Lai H (2013) Using automated processes to generate test items. Educational Measurement: Issues and Practice 32:36–50, DOI 10.1111/emip.12018

Haladyna T, Rodriguez M (2013) Developing and validating test items. Routledge, New York, NY

Hattie J, Timperley H (2007) The power of feedback. Review of Educational Research 88(1):81–112

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

Ho AD, Yu CC (2015) Descriptive statistics for modern test score distributions: Skewness, kurtosis, discreteness, and ceiling effects. Educational and Psychological Measurement 75(3):365–388

Lissitz R, Hou X, Slater S (2012) The contribution of constructed response items to large scale assessment: Measuring and understanding their impact. Journal of Applied Testing Technology 13(3):1–50

Lord FM (1952) The relation of the reliability of multiple-choice tests to the distribution of item difficulties. Psychometrika 17(2):181–194

Moreno R, Martinez R, Muniz J (2015) Guidelines based on validity criteria for the development of multiple choice items. Psicothema 27:388–394, DOI 10.7334/psicothema2015.110

Robinson M, Loch B, Croft T (2015) Student perceptions of screencast feedback on mathematics assessment. International Journal on Research in Undergraduate Mathematics Education (1):363–385, DOI 10.1007/s40753-015-0018-6

Rodriguez M (2011) Item-writing practice and evidence. In: Elliott S, Kettler R, Beddrow P, Kurz A (eds) Handbook of accessible achievement tests for all student: Bridging the gaps between research, practice, and policy, Springer, New York, NY, p 201–206

Salvucci S, Walter E, Conley V, Fink S, Saba M (1997) Measurement error studies at the National Center for Education Statistics. U.S. Department of Education Office of Educational Research and Improvement

Varma S (2006) Preliminary item statistics using point-biseral correlation and p-values. Educational Data Systems Inc, Morgan Hill CA