APPLIED MEASUREMENT IN EDUCATION
2019, VOL. 32, NO. 2, 138–150
https://doi.org/10.1080/08957347.2019.1577249

Partial Credit in Answer-Until-Correct Multiple-Choice Tests Deployed in a Classroom Setting

Aaron D. Slepkov and Alan T. K. Godfrey

Department of Physics & Atronomy, Trent University

ABSTRACT
The answer-until-correct (AUC) method of multiple-choice (MC) testing involves test respondents making selections until the keyed answer is identified. Despite attendant benefits that include improved learning, broad student adoption, and facile administration of partial credit, the use of AUC methods for classroom testing has been extremely limited. This study presents scoring properties and item analysis for 26 AUC university course examinations, administered using a commercial scratch-card response system. Here, we show that beyond the traditional pedagogical advantages of AUC, the availability of partial credit adds psychometric advantages by boosting both the mean item discrimination and overall test-score reliability, when compared to tests scored dichotomously upon initial response. Furthermore we also find a strong correlation between students' initial-response successes and the likelihood that they would obtain partial credit when they make incorrect initial responses. Thus, partial credit is being granted based on partial knowledge that remains latent in traditional MC tests. The fact that these advantages are realized in real-life classroom tests may motivate further expansion of the use of AUC MC tests in higher education.

Introduction

Having been introduced nearly a century ago, multiple-choice (MC) testing tools are becoming ever more widespread at all levels of education. In many contexts, MC testing is increasingly replacing other traditional forms of classroom assessment, such as constructed response (CR) (Nicol, 2007). The principal driver of this change is economic because MC testing is considerably less time- and labor-intensive to score. However, beyond efficiency considerations, MC formats are also being chosen for their demonstrated reliability in measuring student knowledge (Haladyna, 2004). Because MC items require students to select a correct response from a provided list of options, rather than to synthesize an original response, the validity of MC testing has long been questioned (Nicol, 2007). Nonetheless, many studies have established a convincing case for the validity of MC testing (Haladyna, 2004), and this technique is currently the predominant testing format in high-stakes and standardized tests.

Partial-credit scoring is an integral aspect of typical constructed-response testing that is absent from traditional MC formats. Whereas MC scoring is most often dichotomous—correct or incorrect—CR items are typically scored in either continuous or polytomous scales that attempt to ascribe partial credit to the state of partial knowledge displayed by the test taker. For any given test item, nearly all students will possess some level of relevant partial knowledge, and thus scoring schemes that account for such partial knowledge should prove more reliable than ones that do not (Ben-Simon, Budescu, & Nevo,

CONTACT Aaron D. Slepkov ✉ aaronslepkov@trentu.ca Department of Physics & Atronomy, Trent University, 1600 West Bank Drive, Peterborough, ON K9L 0G2, Canada

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/hame.

1997; Hutchinson, 1982). Furthermore, students often view an opportunity to indicate or demonstrate their partial knowledge as more fair than situations where partial knowledge is unaccounted for (DiBattista, Gosse, Sinnige-Egger, Candale, & Sargeson, 2009). Thus, the unavailability of partial credit scoring in traditional MC testing is a major drawback of the technique.

Several strategies for partial-credit scoring of MC tests have been developed, each with attendant benefits and drawbacks (Akeroyd, 1982; Ben-Simon et al., 1997; Bush, 2015; Frary, 1989). Examples of non-computerized techniques that are suitable for classroom testing include (a) subset selection (Bush, 2001; Dressel & Schmid, 1953), (b) confidence scoring (Gardner-Medwin, 1995), (c) elimination testing (Coombs, Milholland, & Womer, 1956), (d) option weighting (Guttman, 1941; Nedelsky, 1954; Serlin & Kaiser, 1978), and (e) option ordering (de Finetti, 1965; Poizner, Nicewander, & Gettys, 1978). At some level, each of these schemes allows for non-dichotomous scoring in an attempt to give credit for students' partial knowledge. For example, the subset selection technique (Bush, 2001) requires students to identify the smallest set of options that they believe contains the keyed (i.e., correct) option. The item is then scored according to how many options are in the selected subset and whether they include the keyed option. Confidence scoring is a variant of this approach, with the student required to distribute a total of 100% of confidence across the set of options. The score on any given item is then simply the confidence value assigned by the student to the keyed option (Gardner-Medwin, 1995). For example, if on a particular item the student indicates an 85% level of confidence that option (b) is the keyed option, while designating 5% each to options (a), (c), and (e) and 0% to option (d), and indeed option (b) is the keyed option, the student receives 85% of the total available marks for the item; if option (c) were the keyed option instead, the student would receive 5% of the total available marks for the item.

Answer-until-correct (Attali, 2011; Epstein et al., 2002; Pressey, 1950) (AUC)—also known as repeated selection (Bush, 2015)—is a unique MC scoring format in which students initially select their single most preferred option, are informed whether they have selected the keyed option, and if not, are then encouraged to select another option, with subsequent confirmatory or corrective feedback being provided until they select the keyed option (DiBattista, 2005). Partial credit is an integral aspect of AUC formats, whereby an item is scored according to how many selections a student had to make en route to selecting the keyed option. The challenge then is to develop response systems that record students' selections as they are made and provide feedback on performance after each selection. Pressey (1926, 1950) described a classroom response system for administering AUC tests/exercises as early as 1926. Currently, a commercially available classroom-ready AUC response form known as the Immediate Feedback Assessment Technique (IFAT) (Epstein et al., 2002) is gaining in popularity in postsecondary education (DiBattista et al., 2009; Persky & Pollack, 2008; Slepkov, 2013). The IFAT response sheet consists of rows of bounded boxes, each covered with an opaque waxy coating similar to that on scratch-off lottery tickets. Each row represents the options for one MC item. For each item, there is only one keyed option, denoted by a small black star in the corresponding option box. Students make their response by scratching the coating off the box of their chosen option. If a black star appears inside the box, the student receives confirmation that the chosen option is correct and proceeds to the next MC item. On the other hand, if no star appears, the student immediately knows that their response is incorrect. The student can then reconsider the question and continue scratching boxes until the star indicating the keyed option is revealed. Partial credit can thus be simply and consistently assigned with the IFAT: full credit is given for items in which the star is revealed with only one box scratched, and partial credit is given for items in which the star is revealed with multiple scratched boxes. Typically, a diminishing amount of partial credit is granted for an increasing number of selections made, with the specific scoring scheme at the discretion of the instructor. Thus, unlike many other partial-credit-granting MC techniques, the IFAT format provides straightforward partial-credit scoring that does not require students either to make introspective judgments (Ben-Simon et al., 1997) or to understand probabilities in order to make optimal selections. Rather, students' optimal test-taking strategy is simply to select the best available option, informed by their knowledge of the subject, and to continue doing so until the keyed option is revealed.

There are several aspects of AUC methods such as the IFAT that make them attractive to instructors and test makers. Pedagogically, the availability of immediate confirmatory/corrective feedback has been demonstrated to promote learning (Epstein et al., 2002), especially of higher-order generalization and knowledge (Attali, 2015; Clariana & Koul, 2005). The partial-credit schemes are straightforward and rational (DiBattista et al., 2009; Slepkov & Shiell, 2014; Slepkov, Vreugdenhil, & Shiell, 2016), engendering a sense of fairness that can be absent in other MC techniques (DiBattista, Mitterer, & Gosse, 2004; Epstein & Brosvic, 2002). Students recognize these advantages (DiBattista, 2006), and 15 years of research have consistently found that they not only embrace the technique but recommend its expanded adoption (DiBattista et al., 2004; Slepkov, 2013). Additionally, by supporting new and sophisticated testing strategies, AUC methods are transforming the way MC testing is used. For example, the attendant benefits of confirmatory/corrective feedback with IFAT has led to the recent development of new MC testing superstructures that build items one upon another, forming integrated testlets designed to assess high-order knowledge typically reserved for CR exams (Shiell & Slepkov, 2015; Slepkov, 2013; Slepkov & Shiell, 2014).

The availability of partial credit inevitably results in increased test scores. However, there is a distinction between higher scores that better represent the state of knowledge of the test taker, and notions of grade inflation—where increases in scores are uncorrelated with students' knowledge. At present, it remains an open question whether the partial credit schemes that can be used in AUC tests are psychometrically or pedagogically justifiable. Early studies suggested that the availability of MC partial credit may indeed boost test-score reliability (Gilman & Ferry, 1972; Hanna, 1975), but these studies were limited by small sample sizes and results were not statistically significant. More recent work has shown that while partial credit in an AUC context improves math test-score reliability (Attali, 2011), such gains were found predominantly for constructed-response AUC, with multiple-choice AUC being "not useful in measurement of partial knowledge" (Attali, Laitusis, & Stone, 2015, 261). Thus, it is important to establish by more direct means that partial-credit schemes in IFAT tests are a discriminating measure of (partial) knowledge. Some recent IFAT studies have demonstrated that partial credit is granted at rates significantly higher than would be expected from repeated selections based on random guessing (DiBattista et al., 2009; Merrel, Cirillo, Schwartz, & Webb, 2015). However, these studies did not demonstrate the existence of a direct correlation between the partial credit earned by students and their level of knowledge of course material, nor did they address the effects of the availability of partial credit on test-score reliability.

Our goal in this study was to examine the effects of granting partial credit with the IFAT on the psychometric properties of real-life as-used classroom tests. In particular, we anticipated that, commensurate with theoretical expectations for polytomous scored tests (McDonald, 1983), the granting of partial credit would lead to increases in the discriminatory power of test items and thus in overall reliability of the test scores. Such a finding would imply a strong link between partial credit and partial knowledge. Thus, we anticipated finding a direct correlation between students' overall knowledge of test material and the extent to which they earned partial credit on MC items that they did not answer correctly on the first attempt. To accomplish these goals, we conduct an analysis of 26 IFAT-administered university-level chemistry and physics exams, comprising a total of ~67,000 MC item responses (485 items) in which partial credit was awarded. We found that the vast majority of the as-given polytomously scored exams were more discriminating and reliable with the allotment of partial credit compared to conditions where the exams were rescored to remove partial credit, and when rescored to simulate the random addition of partial credit after initial incorrect responses. For none of the tests was the dichotomized scoring more reliable.

Methodology

Since 2011, we have used the IFAT to administer midterm and final exams in various introductory physics and chemistry courses at our institution. One of us (ADS) has been closely involved in constructing the majority of these exams via research collaborations with the course instructors, as well as in instructing several of the courses. For this study, we analyzed the test psychometrics of every available exam that had been completed by at least 30 students. In total, 26 IFAT-administered exams are analyzed, as denoted in Table 1. All exams exclusively comprised 5-option items. There is variability in the polytomous scoring used for different exams: 16 of 26 exams employed a [1, 0.5, 0.1, 0, 0] scheme, with full credit given when the keyed option was selected on the first response, half-credit given when the second response was correct, one-tenth credit given on a correct third response, and no credit given for subsequent responses. Other exams employed schemes such as [1, 0.25, 0.125, 0, 0] and [1, 0.4, 0, 0, 0], as denoted in Table 2. The differences in scoring schemes reflect pedagogical and procedural considerations of the various course instructors who deployed the exams. However, as long as significant partial credit is granted for a second selection, the differences in test psychometrics between various rational scoring schemes are expected to be minor (Slepkov et al., 2016). As expected, the choice of scoring scheme most directly affects the average test score, as discussed below.

Item difficulty is defined here as the mean of the item scores. This measure ranges from 1 (all students answer correctly on first attempt) to 0 and decreases with item difficulty. When all items on a test are assigned equal weights, the mean test score for the class can be represented by the mean item difficulty. Thus, the as-given polytomous mean test scores with partial credit are denoted here as $\bar{p}_{\mathrm{pc}}$, for including partial credit.

Item discrimination is a measure of an item's effectiveness at differentiating more knowledgeable from less knowledgeable students (Ebel & Frisbie, 1991). Traditionally, item discrimination is reported via the Pearson product-moment correlation coefficient, r, which for dichotomously scored items is equivalent to the point-biserial correlation coefficient, r_{pbs} (Ebel & Frisbie, 1991). This measure correlates students' scores on a particular item with their total test scores (including the score on the item under consideration), and is thus known as an item-total correlation. As it happens, however, including the item score within the total score tends to increase the magnitude of the item-total correlation (Guilford, 1954), especially when the number of test items is small. Thus, although fairly common in the literature, reporting the item-total correlation as a measure of item discrimination for tests with fewer than ~40 items can be misleading. Because several of the tests examined here had fewer than 40 items, we chose instead to report the more conservative item-excluded correlation, which is independent of the number of test items. This measure of an item's discriminatory power is obtained by correlating students' scores on a particular item with their total test score that excludes that item. The resulting value is referred to as the item-excluded correlation, or as the "corrected item-total correlation," as it is denoted in the statistics package SPSS, for example (Furry & Bacharach, 2014). We computed the item-excluded correlation for each item on our IFAT tests and computed the mean item discrimination for each test, symbolized as $\bar{r}_{\mathrm{pc}}$.

Cronbach's alpha measures a test's internal consistency, and is often reported to represent test–retest reliability (Tavakol & Dennick, 2011). For our polytomously scored IFAT tests, we denote this statistic as $\boldsymbol{\alpha}_{\mathrm{pc}}$.

In order to ascertain the projected psychometrics of our IFAT tests in the absence of partial credit, we artificially dichotomized the data for test items by counting as correct only students' selection of the keyed option on the initial attempt. That is, we used a [1, 0, 0, 0, 0] scoring scheme that awards full credit for a correct initial response and no credit otherwise. We then computed the mean item difficulty ($\bar{p}_{\mathrm{di}}$), the mean item-excluded correlation ($\bar{r}_{\mathrm{di}}$), and the test-score reliability ($\boldsymbol{\alpha}_{\mathrm{di}}$) for each of the dichotomized tests.

Because partial credit invariably contributes an added score above and beyond the baseline dichotomous score, it is difficult to establish a segregated measure of the psychometric effects of partial credit. Thus, starting from our empirical data, we conducted computer simulations to estimate the extent to which random guessing (rather than responses based on partial knowledge) following an initial incorrect selection would affect item discrimination and test reliability. To estimate the effect of a random-guessing approach to option selection, we started with the dichotomized test data and stochastically assigned random partial credit for each item according to the scoring scheme for the test. For example, in a [1, 0.5, 0.1, 0, 0] test, a student selecting an incorrect option on the first try was afforded a 1/4 chance of getting 0.5 marks on the item, and if still unsuccessful, then a 1/3 chance of getting 0.1 marks. This randomization procedure was carried out for each student and for each item on a given test, and $\bar{r}_{\mathrm{di+rand}}$ and $\boldsymbol{\alpha}_{\mathrm{di+rand}}$ were then computed. This randomization procedure was repeated 10,000 times for each test to provide a set of test scoring outcomes in which the awarded partial credit reflected completely random guessing; the mean (±*SD*) for the 10,000 discrimination and reliability values in this set were then computed and are denoted here as $\langle \bar{r}_{\mathrm{di+rand}} \rangle$ and $\langle \boldsymbol{\alpha}_{\mathrm{di+rand}} \rangle$. These simulation algorithms for whole-test analysis were implemented using Python. The remaining item analysis and tests of statistical significance were conducted in Microsoft Excel.

Results

A summary of psychometric measures of the analyzed tests is presented in Tables 1 and 2. An attendant feature of any partial-credit scheme is an increase in test scores. For the 26 as-given tests, the mean test score, averaged across tests and expressed as a percentage, was 61%. With the partial credit removed, the mean dichotomized test score fell by 10 percentage points to 51%. The gain resulting from the granting of partial credit was not constant across tests, but it was well constrained, ranging from 4 to 13 percentage points. As presented in Table 1, in all tests students earned partial credit at a rate exceeding what would be expected from random guessing on selections made following initial incorrect responses. For example, while random guessing alone would be expected to elicit 32% of available partial credit in a [1, 0.5, 0.1, 0, 0] scoring scheme, in all 16 analyzed tests with this scheme, more than 35% of the available partial credit was earned, with the mean being 44% ± 5%. Similar values are found for the other scoring schemes. The "actualized partial credit ratio"—the ratio of obtained partial credit to that expected from random guessing—is greater than 1 for all 26 tests; ranging from 1.2 to 1.9 (see Table 2). Interestingly, there is little correlation (r = 0.07) between this actualized partial credit ratio and the amount of available partial credit (via different scoring schemes). Thus it appears that all of the AUC scoring schemes engender rational selection of repeat responses and lead to increased partial credit in ways that differ from random guessing. These findings are consistent with other reports of IFAT use (DiBattista et al., 2009; Merrel et al., 2015).

Whole-test analysis also provides strong evidence that partial-credit scoring increases the discriminatory power of IFAT items. This can be seen by comparisons of the mean discriminatory power of test items when partial credit is included ($\bar{r}_{\mathrm{pc}}$), excluded ($\bar{r}_{\mathrm{di}}$), and artificially made random $\langle(\bar{r}_{\mathrm{di+rand}})\rangle$. The values for the 26 tests are listed in Table 1 and plotted in Figure 1. With scoring based only on initially correct responses, the mean $\bar{r}_{\mathrm{di}}$ was 0.30 ± 0.06. This value compares favorably with the discriminatory power of other postsecondary classroom exams with dichotomous scoring; for example, DiBattista and Kurzawa (2011) reported a mean discrimination coefficient (i.e., point biserial correlation) of 0.27 ± 0.04 for a diverse set of traditional MC tests. With partial credit being granted, the mean discriminatory power was greater still for 25 of the 26 tests, with the mean $\bar{r}_{\mathrm{pc}}$ being 0.33 ± 0.07. Thus, as shown in Figure 1, the exams were significantly more discriminating with partial credit than without, with a small effect size (paired sample *t*-test; $t(25) = 7.5$, $p < .001$, $d = 0.40$).

To put into perspective the gain in discrimination associated with partial credit, we can consider the decrease in discrimination that would arise if partial credit were to be awarded based on purely random guessing $\langle \bar{r}_{\mathrm{di+rand}} \rangle$. As shown in Figure 1, when we simulate students guessing at random on all repeat selections—thus adding non-discriminating partial credit to a discriminating dichotomous test—$\langle \bar{r}_{\mathrm{di+rand}} \rangle$ is in all cases less than $\bar{r}_{\mathrm{di}}$. A paired-sample *t*-test confirms that this simulated dilution of discrimination is significantly lower than that of the dichotomously scored discrimination, with a small effect size ($t(25) = 5.2$, $p < .001$; $d = 0.23$). It is interesting to note that the inclusion

Table 1. Summary measures of item difficulty, discrimination, and reliability for 26 answer-until-correct, IFAT, tests.

	n	Q	$\bar{p}_{\text{pc}}$	$\bar{p}_{\text{di}}$	$\bar{r}_{\text{pc}}$	$\bar{r}_{\text{di}}$	$\langle \bar{r}_{\text{di+rand}} \rangle \pm$ SD	$r_{\text{pc,di}}$	α_{pc}	α_{di}	$\langle \alpha_{\text{di+rand}} \rangle \pm$ SD
Test 01	205	16	0.59	0.47	0.219	0.187	0.170 ± 0.009	0.954	0.592*	0.540*	0.504 ± 0.018
Test 02	36	16	0.49	0.38	0.232	0.225	0.191 ± 0.024	0.958	0.590	0.583	0.535 ± 0.044
Test 03	63	12	0.6	0.46	0.235	0.229	0.206 ± 0.017	0.960	0.569	0.557	0.516 ± 0.029
Test 04	72	12	0.60	0.48	0.237	0.228	0.207 ± 0.017	0.964	0.579	0.556	0.519 ± 0.30
Test 05	82	8	0.59	0.48	0.241	0.223	0.204 ± 0.019	0.970	0.509	0.484	0.451 ± 0.032
Test 06	189	17	0.57	0.45	0.242	0.203	0.183 ± 0.009	0.956	0.639*	0.570*	0.533 ± 0.017
Test 07	68	12	0.56	0.43	0.259	0.23	0.206 ± 0.017	0.965	0.605*	0.562*	0.522 ± 0.030
Test 08	162	14	0.70	0.63	0.299	0.304	0.304 ± 0.006	0.982	0.679	0.688	0.673 ± 0.08
Test 09	73	8	0.59	0.46	0.317	0.257	0.232 ± 0.021	0.955	0.618*	0.542*	0.504 ± 0.032
Test 10	63	25	0.62	0.5	0.318	0.299	0.274 ± 0.010	0.973	0.791*	0.763*	0.736 ± 0.012
Test 11	158	24	0.55	0.47	0.319	0.298	0.280 ± 0.006	0.980	0.785*	0.761*	0.743 ± 0.006
Test 12	104	21	0.59	0.48	0.321	0.292	0.269 ± 0.009	0.975	0.767*	0.733*	0.704 ± 0.012
Test 13	215	12	0.61	0.53	0.322	0.300	0.284 ± 0.007	0.980	0.685*	0.658*	0.639 ± 0.009
Test 14	36	16	0.6	0.51	0.331	0.28	0.254 ± 0.018	0.970	0.757*	0.687*	0.652 ± 0.026
Test 15	328	32	0.59	0.48	0.332	0.319	0.296 ± 0.004	0.982	0.832*	0.819*	0.798 ± 0.004
Test 16	406	47	0.73	0.64	0.336	0.327	0.309 ± 0.002	0.989	0.874*	0.867*	0.854 ± 0.002
Test 17	51	21	0.49	0.40	0.347	0.294	0.310 ± 0.007	0.995	0.755*	0.735*	0.721 ± 0.008
Test 18	48	20	0.61	0.55	0.366	0.316	0.334 ± 0.007	0.996	0.764*	0.752*	0.738 ± 0.007
Test 19	52	16	0.71	0.62	0.366	0.33	0.310 ± 0.015	0.983	0.763*	0.729*	0.705 ± 0.017
Test 20	78	21	0.67	0.57	0.378	0.367	0.342 ± 0.009	0.964	0.817	0.808	0.786 ± 0.008
Test 21	100	21	0.56	0.44	0.381	0.362	0.334 ± 0.008	0.980	0.820*	0.806*	0.779 ± 0.008
Test 22	97	12	0.62	0.52	0.407	0.405	0.375 ± 0.011	0.977	0.760	0.767	0.739 ± 0.011
Test 23	54	14	0.57	0.51	0.407	0.382	0.392 ± 0.007	0.995	0.771	0.767	0.754 ± 0.006
Test 24	50	16	0.60	0.55	0.423	0.374	0.384 ± 0.007	0.996	0.805*	0.780*	0.768 ± 0.006
Test 25	86	23	0.75	0.71	0.426	0.400	0.413 ± 0.004	0.999	0.844	0.844	0.837 ± 0.003
Test 26	74	28	0.65	0.57	0.438	0.410	0.391 ± 0.007	0.991	0.886*	0.873*	0.866 ± 0.006
Mean			**0.61**	**0.51**	**0.326**	**0.302**	**0.286**	**0.976**	**0.725**	**0.701**	**0.676**
SD			0.06	0.08	0.066	0.065	0.072	0.014	0.106	0.116	0.125

n: Number of students
Q: Number of items
$\bar{p}_{\text{pc}}$: mean polytomously scored item difficulty; mean score on test
$\bar{p}_{\text{di}}$: mean dichotomized test score
$\bar{r}_{\text{pc}}$: mean corrected item-total correlation coefficient for as-scored (polytomous) test; a measure of item discrimination
$\bar{r}_{\text{di}}$: mean corrected item-total correlation coefficient for dichotomized test with partial credit removed
$\langle \bar{r}_{\text{di+rand}} \rangle \pm$ SD: mean item-total correlation coefficient for dichotomized test with added random partial credit *SD*: Standard deviation
$r_{\text{pc,di}}$: correlation between student partial-credit score and dichotomous score on a given test; used for Feldt statistic calculation.
α_{pc}: Cronbach's alpha as a measure of test's reliability, for as-scored (polytomous) test
α_{di}: Cronbach's alpha for test without partial credit (dichotomized)
$\langle \alpha_{\text{di+rand}} \rangle \pm$ SD: Cronbach's alpha for dichotomized test with added random partial credit
*Designates pairs of Cronbach's alpha that differ statistically (Feldt statistic for correlated alphas, $p < .05$)

of partial credit based on partial knowledge tends to boost item discrimination by about the same amount that random guessing decreases it. Taken together, these findings strongly support the notion that partial credit is being granted in a discriminating and valid manner in our AUC tests.

The test reliabilities of the AUC tests are displayed in Figure 2 and listed in Table 1. As a rule of thumb, $\alpha > 0.70$ is considered a benchmark for satisfactory reliability for classroom exams, with values greater than 0.80 being targeted for high-stakes tests (Nunnally, 1978). For our 26 tests, the values of $\boldsymbol{\alpha}_{\text{pc}}$ ranged from 0.51 to 0.89, with 16 tests displaying reliability greater than 0.70. Because Cronbach's alpha scales with the number of test items, the wide range of reliabilities in our tests is at least partially a consequence of the large range of the number of items on the tests. However, regardless of the value of $\boldsymbol{\alpha}_{\text{pc}}$, the granting of partial credit resulted in greater reliability than the application of dichotomized scoring for 17 of 26 exams. For none of the tests did dichotomized scoring prove more reliable. Determination of such differences in reliability was achieved via Feldt's (1980) approach for comparing alphas and is designated (for $p < .05$) in Table 1 and Figure 2. Averaging across exams, the mean $\boldsymbol{\alpha}_{\text{pc}}$ of 0.73 ± 0.11 was significantly greater than the mean $\boldsymbol{\alpha}_{\text{di}}$ of 0.70 ± 0.12, with a small effect size (paired sample t-test; $t(25) = 5.5$, $p < .001$; $d = 0.23$). Furthermore,

Table 2. Scoring schemes and partial credit actualization for 26 answer-until-correct, IFAT, tests.

	Scoring scheme	$\bar{p}_{pc}$	$\bar{p}_{di}$	$\bar{p}_{pc} - \bar{p}_{di}$	$+\bar{p}_{rand}$	Actualized PC ratio
Test 01	[1,0.5,0.1,0,0]	0.59	0.47	0.117	0.079	1.48
Test 02	[1,0.5,0.1,0,0]	0.49	0.38	0.109	0.092	1.18
Test 03	[1,0.5,0.1,0,0]	0.6	0.46	0.132	0.081	1.64
Test 04	[1,0.5,0.1,0,0]	0.60	0.48	0.128	0.078	1.63
Test 05	[1,0.5,0.1,0,0]	0.59	0.48	0.102	0.077	1.32
Test 06	[1,0.5,0.1,0,0]	0.57	0.45	0.126	0.083	1.51
Test 07	[1,0.5,0.1,0,0]	0.56	0.43	0.124	0.085	1.46
Test 08	[1,0.4,0,0,0]	0.70	0.63	0.070	0.037	1.91
Test 09	[1,0.5,0.1,0,0]	0.59	0.46	0.131	0.081	1.61
Test 10	[1,0.5,0.1,0,0]	0.62	0.5	0.120	0.075	1.61
Test 11	[1,0.4,0,0,0]	0.55	0.47	0.086	0.053	1.62
Test 12	[1,0.5,0.125,0,0]	0.59	0.48	0.116	0.082	1.43
Test 13	[1,0.4,0,0,0]	0.61	0.53	0.078	0.047	1.66
Test 14	[1,0.5,0.1,0,0]	0.6	0.51	0.086	0.074	1.17
Test 15	[1,0.5,0.1,0,0]	0.59	0.48	0.112	0.079	1.43
Test 16	[1,0.5,0.1,0,0]	0.73	0.64	0.088	0.054	1.62
Test 17	[1,0.25,0.125,0,0]	0.49	0.40	0.082	0.056	1.46
Test 18	[1,0.25,0.125,0,0]	0.61	0.55	0.058	0.042	1.38
Test 19	[1,0.5,0.1,0,0]	0.71	0.62	0.091	0.057	1.58
Test 20	[1,0.5,0.1,0,0]	0.67	0.57	0.092	0.064	1.45
Test 21	[1,0.5,0.125,0,0]	0.56	0.44	0.125	0.088	1.43
Test 22	[1,0.5,0.1,0,0]	0.62	0.52	0.093	0.072	1.30
Test 23	[1,0.25,0.125,0,0]	0.57	0.51	0.065	0.046	1.41
Test 24	[1,0.25,0.125,0,0]	0.60	0.55	0.054	0.043	1.27
Test 25	[1,0.25,0.125,0,0]	0.75	0.71	0.038	0.027	1.41
Test 26	[1,0.5,0.1,0,0]	0.65	0.57	0.082	0.065	1.26
	Mean	**0.61**	**0.51**	**0.096**	**0.066**	**1.47**
	SD	0.06	0.08	0.026	0.018	0.17

Scoring Scheme: Amount of credit granted for a correct response obtained at the [first try, second try, third try, fourth try, after fourth try].

$\bar{p}_{pc}$: mean polytomously scored item difficulty; mean score on test

$\bar{p}_{di}$: mean dichotomized test score

$\bar{p}_{pc} - \bar{p}_{di}$: mean obtained partial credit

$+\bar{p}_{rand}$: mean expected partial credit obtained for purely guessing (random) repeat responses

Actualized PC ratio: The ratio of obtained partial credit to anticipated partial credit for random guessing

when partial credit is (simulated to be) based on random guessing, the mean test reliability falls to 0.68 ± 0.13, which is significantly less than the reliability of dichotomously scored tests, with a small effect size (paired sample *t*-test; $t(25) = 11.1$, $p < .001$; $d = 0.22$). Again, it is interesting to note that the inclusion of partial credit based on partial knowledge tends to boost test reliability by approximately the same amount as random guessing decreases it.

Evidence for the validity of granting partial credit in the AUC format goes well beyond boosted test scores and whole-test psychometrics. Because partial credit is meant to reflect students' partial knowledge of test material, we would expect the rate at which students obtain partial credit to be directly related to the rate at which they obtain full credit on their first attempt at test items. Indeed, as listed in Table 1, the correlation between the dichotomous scores and the full scores with partial credit is very high, exceeding 0.95 for every tests. Yet a simple correlation between students' dichotomous test scores and the *total* amount of partial credit they earn with AUC scoring would be problematic because the student's dichotomous score caps the available partial credit—that is, the more items the student answers correctly on the initial attempt, the less overall partial credit will be available to them. However, the higher the student's dichotomous score, the more knowledgeable the student, and in general, the greater should be the *proportion* of the *available* partial credit that the student would earn on an AUC test. Figure 3 compares the dichotomized scores and the proportion of available partial credit converted for each student in Test 15. These data suggest significant criterion-related validity evidence for the partial-credit scoring technique, as they convincingly establish a correlation between the ability of students to answer MC items on the first response

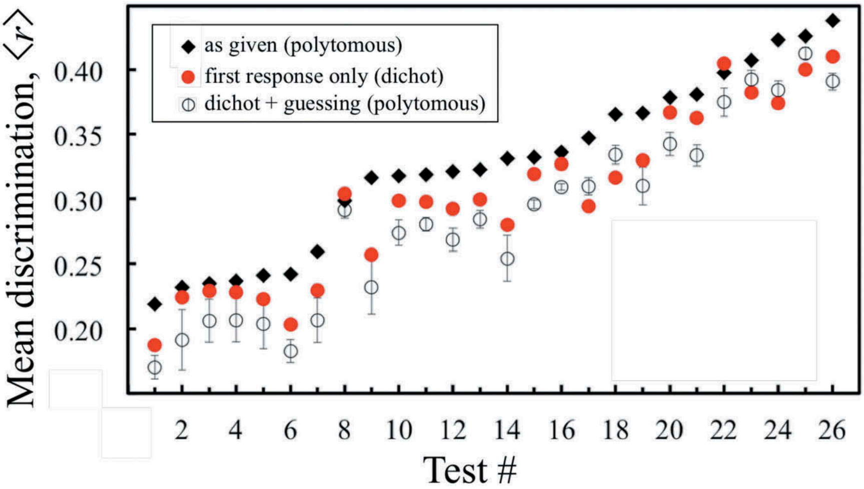

Figure 1. Mean item discrimination measures for each test. The black diamonds represent the IFAT test score with partial credit upon repeat response; the red circles represent the IFAT test scores when only first responses are considered; the empty circles represent artificially simulated IFAT test scores whereby the actual obtained partial credit was removed and replaced with partial credit obtained via purely random guessing. This simulation was run 10,000 times for each test, and the error bars on the empty circles represent that standard deviation of these data.

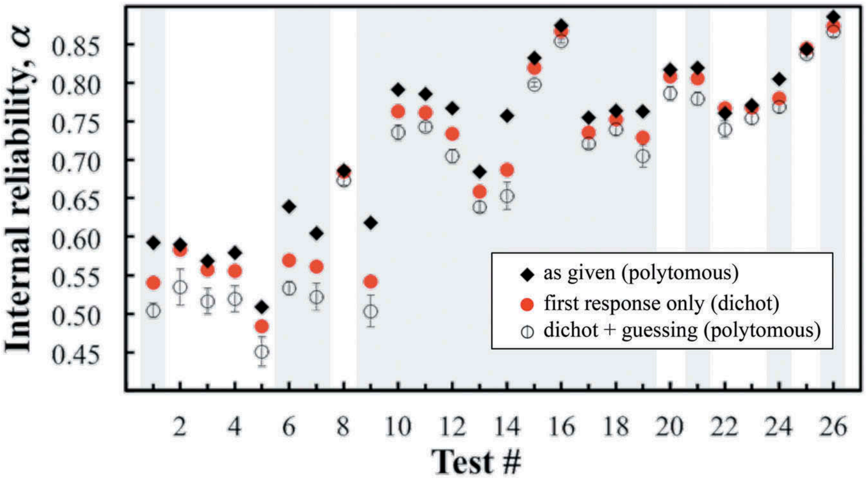

Figure 2. Internal reliability measures (Cronbach's alpha) for each test. The black diamonds represent the IFAT test score with partial credit upon repeat response; the red circles represent the IFAT test scores when only first responses are considered; the empty circles represent artificially simulated IFAT test scores whereby the actual obtained partial credit was removed and replaced with partial credit obtained via purely random guessing. This simulation was run 10,000 times for each test, and the error bars on the empty circles represent the standard deviation of these data. Tests for which the polytomous score is statistically more reliable than for the dichotomous score ($p < .05$) are shaded in grey.

and their ability to overcome initially incorrect responses with a correct subsequent response. Despite substantial variance in the proportion of available partial credit obtained by students, a correlation of 0.52 is found between students' dichotomous scores and the proportion of partial credit they obtained. Note too that the variance in the proportion of obtained partial credit generally increases as the dichotomized score increases, which is not surprising given the inverse correlation between first-response success and the opportunity to obtain partial credit. To reduce the variability in this heteroscedastic data, we binned the data by averaging the obtained partial credit at each

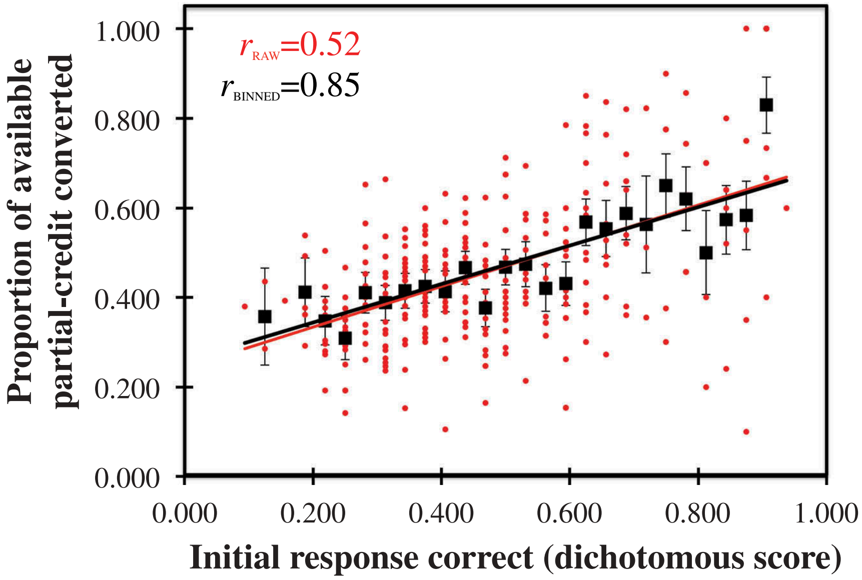

Figure 3. Correlation between dichotomized score and proportion of available partial credit converted for **Test 15**. The red data points represent the complete scatter plot for all students who obtained a dichotomized score lower than 1.00. The raw data from each discrete dichotomized score has been binned and the bin means are represented by black points, with error bars representing the standard error of the mean. Only discrete scores with more than one entry have been binned. Linear correlations are shown via lines of best fit, with correlation coefficients of 0.52 for the raw data and 0.85 for the binned data, respectively, as displayed on the figure.

discrete dichotomized score. We then plotted the bin means and standard errors of the means for the partial-credit scores, including all bins that contained at least two entries. As Figure 3 shows, this yields a much clearer correlation of 0.85 between students' rate of conversion of available partial credit and their dichotomized test scores. Whether raw or binned, both measures of correlation represent a strong effect (Cohen, 1988). It would be ideal to demonstrate this effect in several other tests, but in practice such an intra-test correlational analysis is robust only for tests with a large number of students. Thus, while many of the other tests show this trend (and none show the converse), the pattern of data is not as striking. Besides Test 15, only Test 16 has a sufficient number of students ($N > 300$) for this analysis to be meaningful, and it shows the same trend as Test 15 with similar correlations ($r_{\text{RAW}} = 0.48$ and $r_{\text{BINNED}} = 0.90$) (Slepkov et al., 2016).

Discussion

The pedagogical underpinnings of MC AUC formats such as the IFAT are a significant reason for their increasing adoption. Foremost among these is that immediate corrective/confirmatory feedback has been shown to promote learning. The incentive to obtain such feedback for every item is provided by the opportunities to earn partial credit, with most schemes being based on the number of attempts required to select the keyed response. Thus, the availability of partial credit is an intrinsic component of AUC. Largely because students identify the pedagogical advantages of immediate feedback, and furthermore view scoring schemes that gauge levels of partial knowledge as inherently more fair, they strongly support the expanded use of this technique (DiBattista et al., 2004). It is likely, however, that instructor adoption of the technique has been hindered by worries that the partial credit earned by the students is deficient or even detrimental from a psychometric standpoint. Such worries are perhaps compounded with fears that partial credit therefore acts to inflate grades by unduly boosting test scores. Our results consistently demonstrate that neither worry is justified: For nearly every AUC test analyzed in this study, psychometric measures such as mean item discrimination and test-score reliability are superior with the inclusion of partial credit. This finding will not be surprising to professional psychometricians and quantitative assessment experts, as it has long been known that polytomous scoring can always be made superior to dichotomous scoring. However, empirically establishing that such advantages are realized even in low-stakes classroom examinations should prove useful to a majority of classroom instructors and assessment researchers.

We have demonstrated a clear and strong correlation between a student's propensity to earn partial credit and their dichotomized test score—and therefore their level of knowledge. Ultimately, because the partial credit component in every one of the tests proves discriminating, such partial credit is rational and does not represent score inflation. While it is true that availability of partial credit inevitably increases test scores, the practicality of providing such partial credit depends on the extent to which the baseline (dichotomous) score increases. The physics and chemistry tests analyzed here were all sufficiently challenging to make viable the addition of partial credit. With as-given test scores ranging from 0.49 to 0.75, we faced little concern that the availability of partial credit overinflated these summative tests. Negative-scoring schemes that are designed to eliminate positive score gains from random guessing (e.g., Bush, 2015) are inconsistent with the objectives of AUC in that they remove the assurance that students will eventually discover the keyed option for each item. Regardless, as our results demonstrate, there is no evidence that the availability of partial credit is *inflating* test scores. Thus, there is a strong case to be made that the use of AUC test formats that include partial credit schemes, such as the IFAT, is also psychometrically motivated.

The improvement in item discrimination and test reliability that arises from the availability of partial credit with AUC is statistically significant and extremely consistent. This finding supports the conclusion of Attali that "the difference between initial and revised scores lies in more precise trait measurement and not in measurement of a different trait" (2011, p. 478). The effect sizes of the boost in discrimination and reliability that arise from the granting of partial credit may appear to be modest, but a comparison to the maximum effect sizes that can be expected provides strong support for concluding that the effects of partial credit are relatively large in this study. For example, while typical effect sizes, via Cohen's d statistic, range from zero (for no effect), through 1 (a large and notable effect), to infinity (for maximum effect), in the case of test scores the availability of partial credit is bounded below by the dichotomized score and above by a perfect score, and thus there is a relatively low upper-bound on the possible value of d. Furthermore, a realistic maximal effect size would arise from the case that partial credit is perfectly correlated with a student's baseline knowledge. We estimate that if partial credit were to be granted at the exact same rate as a student's dichotomous score, that the effect size for that case would be capped below a value of $d = 1.0$ for test-score reliabilities and below $d = 2.5$ for mean item discriminations. Considering that the obtained mean gains in test-score reliability with AUC show an effect size of $d = 0.23$, and gains in item discrimination show $d = 0.40$, our experimental effect sizes must be reinterpreted to suggest that the availability of partial credit via the IFAT represents positive, consistent, significant, and strong effects on test psychometrics.

This study provides consistent and strong evidence of the effectiveness of the IFAT format for providing discriminating partial credit. However, the comparisons between test scores with and without partial credit necessarily required a post-hoc analysis. These post-hoc dichotomized scores might not faithfully represent how students approach a traditional MC test when they know that partial credit will not be available. Thus, in a future study it would be desirable to compare tests with and without the availability of partial credit where both tests utilize the same AUC format, but with one cohort knowing in advance that they will not earn partial credit. This is important because it is possible that foreknowledge of the availability of partial credit upon repeat selection affects the behavior of the test taker, perhaps making their selections more compulsive and less mindful (Attali, Laitusis, & Stone, 2016). Conducting such a study with traditional MC items might be difficult due to the need to maintain the incentive for students to continue making selections until the keyed option is obtained. If students obtain credit only when their first selection is correct, they might discount the benefits of corrective feedback that come from finding the keyed response, because of the added time required to do so. However, this difficulty might be mitigated with the use of integrated testlets (Slepkov & Shiell, 2014), wherein items are sequentially inter-dependent, therefore providing strong non-partial-credit-based incentive to identify the keyed option because the information provided might be useful for answering subsequent items.

A purported pedagogical advantage of AUC formats is that immediate corrective feedback acts as an intervention with strong aspects of formative learning. Rather than allowing student misinformation to crystallize, corrective feedback forces students to revise their way of thinking *en route* to making a correct response. This effect is related to the fact that there are important but subtle differences between Option Ordering and Repeated Selection (as in AUC). For example, consider the case of the AUC where a student fails to identify the keyed option on the first response and is faced with the opportunity to identify it on the second try. Ideally, the student would assess their initial selection and with the knowledge that this selection was incorrect then reassess which remaining option is most likely to be correct with the conditional probability that their best first choice was incorrect. Because it is likely that their initial ordering of all options was based on the same flawed thinking that led to an incorrect initial selection, a correction to that thinking may well lead to a reordering of the remaining options. This is key to the targeted advantage of AUC. On the other hand, a "mindless" automatic selection of their initially assessed second best choice might short-circuit these key aspects of corrective feedback. In that case, Repeated Selection would largely mimic Option Ordering, losing much of its pedagogical advantage. While this study establishes a clear link between sequential selections and student knowledge, it is not able to identify whether students react to an incorrect initial response with secondary responses based on their initial (mis)conceptions, or whether the corrective feedback of the IFAT results in a formative reassessment of such (mis)conceptions. Such disambiguation may be resolved with future cognitive-lab studies that record think-aloud responses (Leighton, 2017) to AUC tests, and can thus track the level of re-evaluation and revision that takes place after an initial incorrect response.

Summary

We have presented analysis of 26 university chemistry and physics multiple-choice classroom exams, all of which utilized an answer-until-correct response format that granted partial credit based on the number of repeated responses students required before selecting the keyed option. The classroom tests were administered using the IFAT, a commercial AUC scratch-card response system. Past research has established numerous advantages of AUC multiple-choice testing, not the least of which are the promotion of learning and strong student adoption. Without the use of negative-scoring scheme, test scores are invariably higher in AUC formats that grant partial credit based on the number of selections made by students en route to a correct response. We find that, on average, test scores increase by 10 percentage points, from a mean score of 51% for initial response scoring to that of 61% with repeat selection utilizing [1, 0.5, 0.1 (or 0.125), 0, 0] scoring. With test scores in the low 60s, there is little concern that the availability of partial credit is untenably inflating test scores. On the other hand, consistent with well-known principles that polytomous scoring is superior psychometrically to dichotomous scoring, we find that the availability of partial credit significantly boosts both the mean item discrimination and overall test-score reliability in AUC tests. Furthermore, within the tests with the largest number of students we also find a strong correlation between students' initial-response successes and the likelihood that they would obtain partial credit when they make incorrect initial responses. Such correlations represent a large effect that suggests that partial credit is being granted based on partial knowledge that may typically be unmeasured in traditional multiple-choice tests. Thus, in addition significant pedagogical motivations, there are strong psychometric reasons to motivate further expansion of AUC MC testing in higher education. Beyond the IFAT, there are now numerous computer-administered AUC tools ("Learning Catalytics," 2015; "QuizSlides," 2015) that both simplify the administration of AUC and provide new opportunities for reaping the unique pedagogical benefits of this powerful multiple-choice technique.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Aaron D. Slepkov http://orcid.org/0000-0001-5699-1403

References

Akeroyd, F. M. (1982). Progress in multiple-choice scoring methods 1977–81. *Journal of Further and Higher Education*, *6*, 87–90.

Attali, Y. (2011). Immediate feedback and opportunity to revise answers: Application of a graded response IRT model. *Applied Psychological Measurement*, *35*, 472–479. doi:10.1177/0146621610381755

Attali, Y. (2015). Effects of multiple-try feedback and question type during mathematics problem solving on performance in similar problems. *Computers & Education*, *86*, 260–267. doi:10.1016/j.compedu.2015.08.011

Attali, Y., Laitusis, C., & Stone, E. (2015). Effects of multiple-try feedback and question type during mathematics problem solving on performance in similar problems. *Computers & Education*, *86*, 260–267. doi:10.1016/j.compedu.2015.08.01

Attali, Y., Laitusis, C., & Stone, E. (2016). Differences in reaction to immediate feedback and opportunity to revise answers for multiple-choice and open-ended questions. *Educational and Psychological Measurement*, *76*, 787–802. doi:10.1177/0013164415612548

Ben-Simon, A., Budescu, D. V., & Nevo, B. (1997). A comparative study of measures of partial knowledge in multiple-choice tests. *Applied Psychological Measurement*, *21*, 65–88. doi:10.1177/0146621697211006

Bush, M. (2001). A multiple choice test that rewards partial knowledge. *Journal of Further and Higher Education*, *25*, 157–163. doi:10.1080/03098770120050828

Bush, M. (2015). Reducing the need for guesswork in multiple-choice tests. *Assessment & Evaluation in Higher Education*, *40*, 218–231. doi:10.1080/02602938.2014.902192

Clariana, R. B., & Koul, R. (2005). Multiple-try feedback and higher-order learning outcomes. International. *Journal of Instructional Media*, *32*, 239–245.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Assoc.

Coombs, C. H., Milholland, J. E., & Womer, F. B. (1956). The assessment of partial knowledge. *Education and Psychological Measurement*, *16*, 13–37. doi:10.1177/001316445601600102

de Finetti, B. (1965). Methods for discriminating levels of partial knowledge concerning a test item. *British Journal of Mathematical and Statistical Psychology*, *18*, 87–123. doi:10.1111/bmsp.1965.18.issue-1

DiBattista, D. (2005). The immediate feedback assessment technique: A learner-centered multiple-choice response form. *The Canadian Journal of Higher Education*, *35*, 111–131.

DiBattista, D. (2006). Test anxiety and the immediate feedback assessment technique. *The Journal of Experimental Education*, *74*, 311–327. doi:10.3200/JEXE.74.4.311-328

DiBattista, D., Gosse, L., Sinnige-Egger, J.-A., Candale, B., & Sargeson, K. (2009). Grading scheme, test difficulty, and the immediate feedback assessment technique. *Journal of Experimental Education*, *77*, 311–336. doi:10.3200/JEXE.77.4.311-338

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

DiBattista, D., Mitterer, J. O., & Gosse, L. (2004). Acceptance by undergraduates of the immediate feedback assessment technique for multiple-choice testing. *Teaching in Higher Education*, *9*, 17–28. doi:10.1080/1356251032000155803

Dressel, P. L., & Schmid, J. (1953). Some modifications of the multiple-choice item. *Educational and Psychological Measurement*, *13*, 574–595. doi:10.1177/001316445301300404

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed., pp. 231–232). Englewood Cliffs, CA: Prentice-Hall.

Epstein, M. L., & Brosvic, G. M. (2002). Students prefer the immediate feedback assessment technique. *Psychological Reports*, *90*, 1136–1138. doi:10.2466/pr0.2002.90.3c.1136

Epstein, M. L., Lazarus, A. D., Calvano, T. B., Matthews, K. A., Hendel, R. A., Epstein, B. B., & Brosvic, G. M. (2002). Immediate feedback assessment technique promotes learning and corrects inaccurate first responses. *The Psychological Record*, *52*, 187–201. doi:10.1007/BF03395423

Feldt, L. S. (1980). A test of the hypothesis that Cronbach's alpha reliability coefficient is the same for two tests administered to the same sample. *Psychometrika*, *45*, 99–105. doi:10.1007/BF02293600

Frary, R. B. (1989). Partial-credit scoring methods for multiple-choice tests. *Applied Measurement in Education*, *2*, 79–96. doi:10.1207/s15324818ame0201_5

Furry, R. M., & Bacharach, V. R. (2014). *Psychometrics: An introduction* (2nd ed., pp. 180). Thousand Oaks, CA: SAGE Publications.

Gardner-Medwin, A. R. (1995). Confidence assessment in the teaching of basic science. *Association for Learning Technology Journal*, *3*, 80–85. doi:10.3402/rlt.v3i1.9597

Gilman, D. A., & Ferry, P. (1972). Increasing test reliability through self-scoring procedures. *Journal of Educational Measurement*, *9*, 205–207. doi:10.1111/jedm.1972.9.issue-3

Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). New York, NY: McGraw-Hill.

Guttman, L. (1941). An outline of the statistical theory of prediction. In P. Horst (Ed.), *The prediction of personal adjustment* (pp. 253–3 11). New York, NY: Social Science Research Council.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Hanna, G. S. (1975). Incremental reliability and validity of multiple-choice tests with an answer-until-correct procedure. *Journal of Educational Measurement*, *12*, 175–178. doi:10.1111/jedm.1975.12.issue-3

Hutchinson, T. P. (1982). Some theories of performance in multiple choice tests, and their implications for variants of the task. *British Journal of Mathematical and Statistical Psychology*, *35*, 71–89. doi:10.1111/bmsp.1982.35.issue-1

Learning Catalytics is a commercial online physics instruction platform that incorporates various testing platforms including AUC capabilities. (2015, October 14). Retrieved from https://learningcatalytics.com/

Leighton, J. P. (2017). *Using think-aloud interviews and cognitive labs in educational research*. New York, NY: Oxford University Press.

McDonald, R. P. (1983). Alternative weights and invariant parameters in optimal scaling. *Psychometrika*, *48*, 377–391. doi:10.1007/BF02293682

Merrel, J. D., Cirillo, P. F., Schwartz, P. M., & Webb, J. A. (2015). Multiple-choice testing using immediate feedback—assessment technique (IF AT®) forms: Second-chance guessing vs. Second-chance learning? *Higher Education Studies*, *5*, 50–55.

Nedelsky, L. (1954). Ability to avoid gross error as a measure of achievement. *Educational and Psychological Measurement*, *14*, 459–472. doi:10.1177/001316445401400303

Nicol, D. (2007). E–assessment by design: Using multiple–choice tests to good effect. *Journal of Further and Higher Education*, *31*, 53–64. doi:10.1080/03098770601167922

Nunnally, J. C. (1978). *Psychometric Theory* (2nd ed.). New York, NY: McGraw-Hill.

Persky, A. M., & Pollack, G. M. (2008). Using answer-until-correct examinations to provide immediate feedback to students in a pharmacokinetics course. *American Journal of Pharmaceutical Education*, *72*(4), 83. doi:10.5688/aj720483

Poizner, S. B., Nicewander, W. A., & Gettys, C. F. (1978). Alternative response and scoring methods for multiple-choice items: An empirical study of probabilistic and ordinal response modes. *Applied Psychological Measurement*, *2*, 83–96. doi:10.1177/014662167800200109

Pressey, S. L. (1926). A simple apparatus which gives tests and scores and teaches. *School and Society*, *23*(586), 373–376.

Pressey, S. L. (1950). Development and appraisal of devices providing immediate automatic scoring of objective tests and concomitant self-instruction. *Journal of Psychology*, *29*, 417–447. doi:10.1080/00223980.1950.9916043

QuizSlides is a web-based application that allows creation of interactive tests, including answer- until-correct multiple-choice items. (2015, October 14). Retrieved from https://quizslides.com/features#Overview

Serlin, R. C., & Kaiser, H. F. (1978). A method for increasing the reliability of a short multiple-choice test. *Educational and Psychological Measurement*, *38*, 337–340. doi:10.1177/001316447803800214

Shiell, R. C., & Slepkov, A. D. (2015). Integrated testlets: A new form of expert-student collaborative testing. *Collected Essays in Teaching and Learning (CELT)*, *8*, 201–210. doi:10.22329/celt.v8i0.4244

Slepkov, A. D. (2013). Integrated testlets and the immediate feedback assessment technique. *American Journal of Physics*, *81*, 782–791. doi:10.1119/1.4820241

Slepkov, A. D., & Shiell, R. C. (2014). Comparison of integrated testlet and constructed- response question formats. *Physical Reviews Special Topics—Physics Education Research*, *10*, 020120. doi:10.1103/PhysRevSTPER.10.020120

Slepkov, A. D., Vreugdenhil, A. J., & Shiell, R. C. (2016). Score increase and partial-credit validity when administering multiple-choice tests using an answer-until-correct format. *Journal of Chemical Education*, *93*, 1839–1846. doi:10.1021/acs.jchemed.6b00028

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, *2*, 53–55. doi:10.5116/ijme.4dfb.8dfd