JOURNAL OF CHEMICAL EDUCATION

Article

pubs.acs.org/jchemeduc

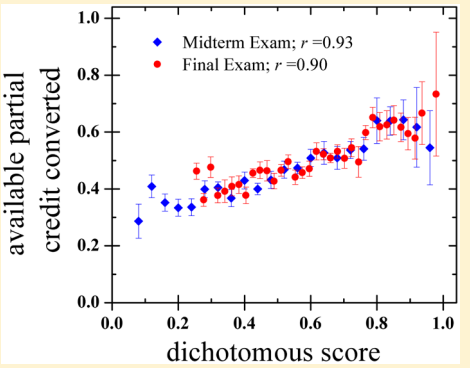
Score Increase and Partial-Credit Validity When Administering Multiple-Choice Tests Using an Answer-Until-Correct Format

Aaron D. Slepkov,*[,†] Andrew J. Vreugdenhil,[‡] and Ralph C. Shiell[†]

[†]Department of Physics & Astronomy, Trent University, Peterborough, Ontario K9L 0G2, Canada

[‡]Department of Chemistry, Trent University, Peterborough, Ontario K9J 0G2, Canada

ABSTRACT: There are numerous benefits to answer-until-correct (AUC) approaches to multiple-choice testing, not the least of which is the straightforward allotment of partial credit. However, the benefits of granting partial credit can be tempered by the inevitable increase in test scores and by fears that such increases are further contaminated by a large random guessing component. We have measured the effects of using the immediate feedback assessment technique (IF-AT), a commercially available AUC response system, on the scores of a typical first-year chemistry multiple-choice test. We find that with a particular commonly used scoring scheme the test scores from IF-AT deployment are 6−7 percentage points higher than from Scantron deployment. This amount is less than that suggested by previous studies, where the mark increase was calculated in a purely *post hoc* manner and thus neglected affective changes of students' behavior associated with the IF-AT technique. Furthermore, we have strong evidence that partial credit is awarded in a highly rational manner in accordance with the students' level of understanding.



KEYWORDS: *Testing/Assessment, High School/Introductory Chemistry, First-Year Undergraduate/General, Curriculum, Student-Centered Learning, Problem Solving/Decision Making*

FEATURE: Chemical Education Research

INTRODUCTION

Multiple-choice (MC) testing methods are widely used in formal assessments within many large introductory chemistry courses and other higher-education STEM courses.[1−3] While this use is principally motivated economically, as MC questions require considerably less time and labor to score, MC methods have nonetheless been demonstrated by many researchers to be valid and highly reliable tools for measuring student knowledge.[4,5] Yet, despite their ubiquity, traditional MC tests still lack key pedagogical attributes such as availability of partial-credit, multistep question structuring, and immediate feedback. Beyond traditional "Scantron"[6]-type formats that score MC questions dichotomously as either right or wrong, numerous alternative formats and scoring schemes have been devised over the past century to allow the granting of partial credit in an effort to better gauge the students' level of partial knowledge,[7−10] often improving test reliability.[9,11,12]

Examples of such approaches include manipulation of the choices given to students so that options contain different combinations of primary responses, only some of which are true (complex multiple-choice, type K, true-false or type X, and multiple-response formats);[13] manipulation of the stems by asking students for predictive or evaluative assessments of a scenario rather than simply recounting knowledge;[13] confidence or probability weighting of options,[8] and the "multiple response format" in which multiple stages are created within each multiple-choice item, with scores weighted according to whether the reasoning is correct.[14] All these schemes suffer from, as Ben Simon et al. relate,[9] the challenge of (mis-)interpreting the intention and state of knowledge of the student.

A particularly simple MC testing method for granting partial credit without needing to interpret student knowledge or intentions is known as answer-until-correct (AUC).[15,16] The AUC method allows for straightforward administering of partial credit based on the number of sequential selections required to identify the correct answer. In contrast to methods that require students and/or instructors to assess the relative merits of ultimately incorrect options,[7,17−20] with AUC the partial credit is invariably tied to selection of the fully correct response. Thus, AUC methods, by virtue of providing immediate confirmation of the correct answer, reduce the risk of propagating or crystallizing student misinformation.[21−23] Deployment of an AUC approach requires a technology that provides an indication of the veracity of the response after each selection. As early as 1926, Pressey developed a mechanical AUC system for classroom use.[24] Since then, commercial techniques that utilize erasable carbon paper were briefly used, and more recently, lottery-style scratch cards have been developed.[25] Digital computer applications can offer agile and customizable

Received: January 13, 2016
Revised: August 12, 2016

DOI: 10.1021/acs.jchemed.6b00028
J. Chem. Educ. XXXX, XXX, XXX−XXX

administration of AUC tests,[26] and their usage is poised for rapid growth, but they have been slow in adoption as most course testing is still administered in-class by pen-and-paper means.

One scratch-card product that has been gaining in popularity is the immediate feedback assessment technique, or IF-AT.[16,27,28] The IF-AT response sheet consists of rows of bounded boxes, each covered with an opaque waxy silver coating. Each row represents the options for one MC item. For every item there is only one keyed answer, represented by a small black star within the corresponding box. Students make a selection by scratching the coating off the box that represents their chosen option. If the scratched box reveals the black star, the student knows that they have correctly selected the keyed response. Alternatively, if no star appears, the student immediately knows that their chosen option is incorrect. The student can then reconsider the question and continue scratching boxes until the star indicating the keyed option is revealed or until scoring opportunities are exhausted. Typically students are encouraged to reveal the keyed response regardless of scoring opportunities, with no disadvantage in doing so.

There are three aspects of AUC methods such as the IF-AT that make them attractive to instructors and test makers. First, the availability of immediate confirmatory/corrective feedback has been demonstrated to promote learning,[16] especially of higher-order generalization and knowledge.[29,30] Second, the partial-credit schemes are straightforward and rational,[31,32] engendering a sense of fairness that can be absent in other MC techniques.[33,34] Indeed, 15 years of research has consistently found that students approve of the IF-AT and recommend its expanded adoption.[33,35] Third, the attendant benefit of confirmatory feedback in AUC formats has allowed for the development of new MC testing superstructures wherein items build one upon another. Known as integrated testlets (ITs), the use of these structures in STEM disciplines is motivated by the desire to assess higher-order knowledge that is typically reserved for traditional constructed-response questions.[31,35] The development of ITs has specifically targeted STEM disciplines, where the assessment of integrated conceptual and procedural understanding is of prime importance but difficult to attain with the traditional MC format.[36]

When assessing higher-order thinking, the absence of a simple means of rewarding incomplete understanding is a particular drawback of traditional MC tests. Additionally, concurrent with the desire for more nuanced scoring of MC questions is a pervading aversion against rewarding students for random guessing.[8,37−39] These two considerations are in conflict in the AUC approach because partial credit is granted via repeated selections, which necessarily provide added opportunities for random guessing with continually improving odds. In other MC formats, a penalty for random guessing is sometimes instituted in the form of "negative scoring" schemes,[8,37,40] but because in AUC students are required to make selections until they discover the keyed response, negative scoring is incompatible with the IF-AT.[32] Thus, to address questions regarding both grade inflation and possible increased guessing due to repeated response and the concomitant awarding of partial credit, an understanding of the effects of using the IF-AT on test scores is needed.

Past research on the operation of the IF-AT establishes that with various scoring schemes similar to that used here, the allotment of partial credit increases test scores by approximately 8−13 percentage points.[35,41−43] In these studies an *estimate* of the effects of partial credit is made *post hoc* by rescoring AUC tests as traditional first-selection dichotomous tests. This measurement therefore neglects any *affective behavior* changes due to students approaching IF-AT tests differently than they do Scantron tests. Such effects could be significant: For example, instructors who adopt IF-AT/AUC tests report concerns regarding a possible increased carelessness in initial selections, arising from the perceived safety-net of partial credit on repeated selection.[30,31] Additionally, "lottery-style" scratch cards may engender a greater sense of required "luck" than do traditional response cards, thereby increasing the likelihood of student guessing. Finally, the immediate feedback aspects of the technique mean that students are confronted by their lack of knowledge during the assessment, and this has led to questions regarding student anxiety and its effect on test scores through different student behaviors.[44] Thus, a comparison of results from the *post hoc* dichotomization of IF-AT tests with those from Scantron tests may not accurately reflect the differences in test scores due to repeated selection alone. In this paper we report on a two-year study in a large introductory chemistry course that compares one midterm delivered through Scantron with a concept-equivalent midterm delivered through IF-AT, and one final delivered through Scantron with the *identical* final delivered through IF-AT.

The following research questions guided this investigation: (1) What effect on test scores does moving from Scantron to IF-AT have in a typical introductory chemistry course final exam? (2) What evidence is there that partial credit granted via IF-AT for selections beyond the first response is distributed in a valid manner, and are such responses more or less random than the first selection? (3) What evidence is there that a dichotomously scored IF-AT exam is answered differently than a traditional multiple-choice exam such as Scantron-type, and is there a score penalty or boost to any such affective behavior?

The approach described herein allowed us to measure the differential effects of using the IF-AT response system over Scantron on the test scores in a typical high-enrollment undergraduate STEM class, and to furthermore assess whether the partial credit is granted rationally via IF-AT. We find that, with a typical scoring scheme, IF-AT scores average approximately 6−7 percentage points higher than those of equivalent dichotomous MC tests. Additionally, partial credit is indeed earned in a consistent and highly discriminating manner. Finally, an analysis of the number of options in items further suggests that the effects of random guessing are largely negligible in well-constructed IF-AT tests.

METHODOLOGY

We compare midterm and final exam scores obtained from two consecutive year offerings of a second-term introductory chemistry course (Chemistry-II), where, in the first year, the midterm and final exam delivery was through Scantron, and in the second year delivery of these was through IF-AT. A first-term course, Chemistry-I, precedes this second-term course, and for both years in Chemistry-I the midterm and final exam delivery was through Scantron. The curriculum for the two courses is largely distinct, with Chemistry-I covering reactions, atomic structure, bonding, intermolecular interactions, and chemical equilibria. Chemistry-II covers thermodynamics, electrochemistry, kinetics, and an introduction to organic, biochemistry, and transition metal chemistry. The courses were

DOI: 10.1021/acs.jchemed.6b00028
J. Chem. Educ. XXXX, XXX, XXX−XXX

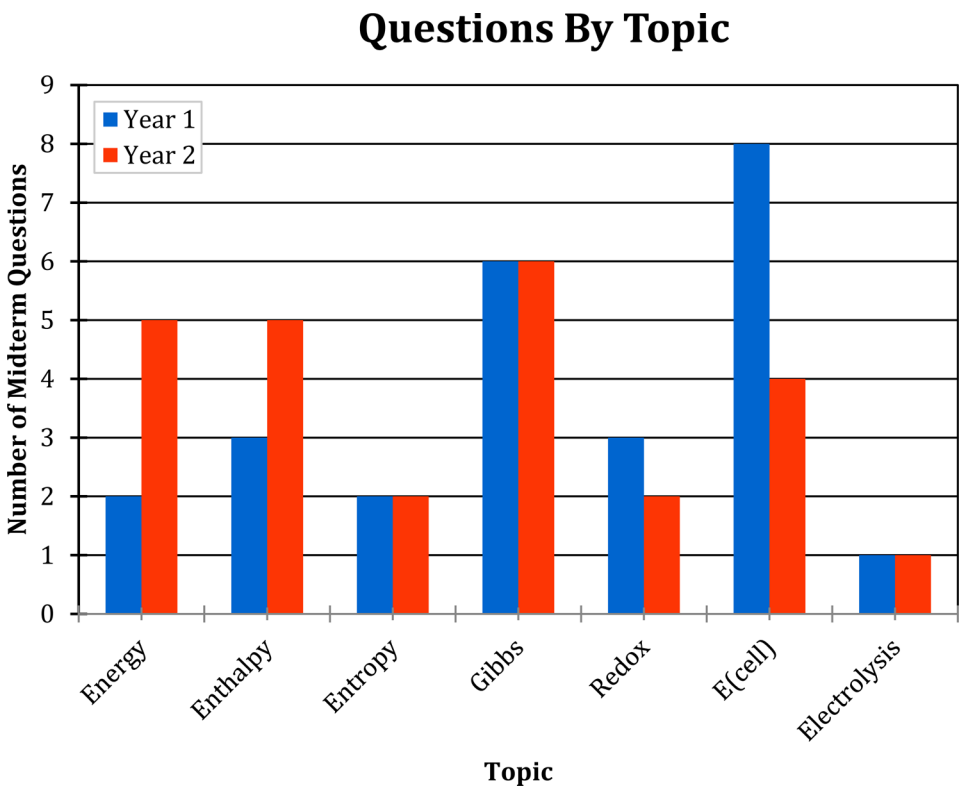


Figure 1. Number of questions on the Chemistry-II midterms in Year 1 and Year 2, by concept.

delivered at a small, primarily undergraduate Canadian university.

Both courses are required for chemistry, biology, and forensic science majors. The student population is broad in interests and anticipated majors. Thus, the results from this investigation are relevant to a range of STEM disciplines.

The same instructor taught Chemistry-I in the fall terms of Year 1 and Year 2 of the study. In the Winter terms, Chemistry-II was taught by different instructors in Year 1 and Year 2, both of whom have taught the course several times before, working from common lecture materials. These two offerings were closely aligned by lecture content, course textbook, instructor consultation, laboratories, and exams.

For Chemistry-I, midterm and final exams for both Year 1 and Year 2 of the study comprised MC questions from a major publisher's testbank.[45] For Chemistry-II, in Year 1 the midterm and final exams also utilized testbank questions, with the midterm comprising 25 questions and the final exam comprising 50 questions. For Chemistry-II in Year 2, the midterm exam comprised 25 questions from the same testbank, without particular consideration of the questions selected for the Chemistry-II midterm in Year 1. By serendipity, five of the 25 questions on the Chemistry-II midterm in Year 1 appeared on the Year 2 midterm. The Chemistry-II midterms for Year 1 and Year 2 of the study covered the same course content, comprising thermodynamics and electrochemistry from three chapters in the textbook common to the Year 1 and Year 2 offerings of Chemistry-II. Figure 1 shows a breakdown of the number of questions in each of the midterms by concept. A difficulty score on a simple three-point-scale (easy 1, moderate 2, and difficult 3) is included by the publisher for each question in the testbank. The average difficulties for the questions on the midterms in Year 1 and Year 2 were the same (2.28). The average difficulties for the thermodynamics questions were 2.31 and 2.28 in Year 1 and Year 2, respectively, and those for the electrochemistry questions were 2.25 and 2.29 in Year 1 and 2.

The final exams for Chemistry-II for Year 1 and Year 2 used the identical set of questions, but in Year 2 they were deployed using IF-AT. Thus, Year 1 and Year 2 midterms in Chemistry-II are said to be "concept-equivalent" because they cover the same course topics without necessarily using identical questions, while the final exams are termed identical in that they used the same questions year-over-year with a (near-) identical item order. The scoring scheme used with the IF-AT in Chemistry-II Year 2 midterm and final exams rewarded full credit for revealing the keyed response in the first selection, 1/2 credit for revealing it in the second selection, 1/10 credit for revealing it in the third response, and no credit afterward. This scheme is denoted [1, 0.5, 0.1, 0, 0]. A representative example of an IF-AT card, with scoring scheme, is shown in Figure 2. Because the IF-AT cards are prekeyed, ensuring the matching of items and requisite key required a different option ordering in each exam version. For test security reasons, two versions each of the midterm and final exams were deployed, each comprising identical questions but with quasirandomized option ordering. Aside from this reordering, the two versions were identical. It was discovered during deployment of the Chemistry-II Year 2 final exam that two IF-AT items were miskeyed, and later found that one item was not identical between the two years' versions. Thus, for analysis we removed 3 of the 50 final exam questions, leaving 47 matched questions in the 2013 and 2014 Chemistry-II final exam for analysis. Of these, 13 items were of the 4-option type, and 34 items were of the 5-option type.[46]

The numbers of students completing the final exam in each of the respective courses follow: 2012 Chemistry-I, 385; 2013 Chemistry-I, 453; 2013 Chemistry-II, 346; 2014 Chemistry-II, 406. Not all of the students who complete Chemistry-I proceed to take Chemistry-II in the immediately following term.

DOI: 10.1021/acs.jchemed.6b00028
J. Chem. Educ. XXXX, XXX, XXX−XXX

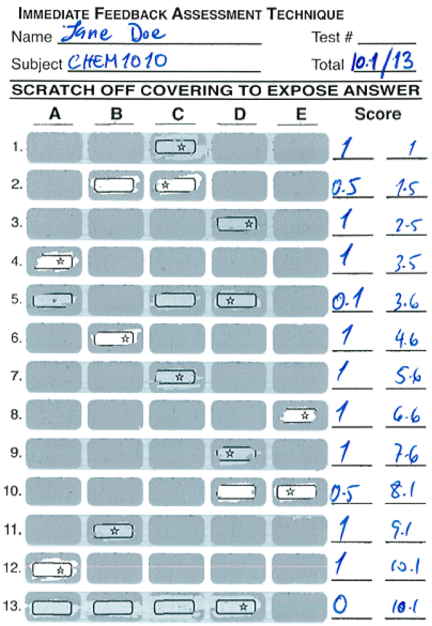


Figure 2. First 13 rows (with mock responses) of an exemplar IF-AT scratch card. Students make selections for a given item (horizontal row) until the keyed response is revealed by the presence of a star within a bounding box. Partial credit is granted based on the number of selections required to reveal the star. The displayed card may be self-scored, as shown using a [1, 0.5, 0.1, 0,0] scoring scheme.

Ultimately, 314 students completed both Chemistry-I and Chemistry-II in Year 1 of the study, and 367 students completed both Chemistry-I and Chemistry-II in Year 2 of the study. This retrospective investigation required no interventions beyond those usually adopted within a program review of a typical iteration of a course. We report here on results obtained in the process of pedagogical review and educational tool assessment, and thus REB/IRB review was not required.

RESULTS AND DISCUSSION

Because the Scantron and IF-AT formats are being compared among cohorts from different years, it is important to calibrate these cohorts' baseline abilities. Aiding this goal is the fact that both populations received instruction in Chemistry-I from the same person and with essentially identical syllabuses. The fact that the Chemistry-I final exams in each of Year 1 and Year 2 comprised the same number of MC items from the same testbank suggests that these constitute a good baseline measure of the incoming abilities of the Year 1 and Year 2 Chemistry-II populations. Indeed, we find that the mean Chemistry-I final exam scores, pared down only to those students who are included in the Chemistry-II Scantron/IF-AT comparison, to be nearly identical, at 57.5% ± 14.1% (M ± SD) and 57.9% ± 15.5%, respectively, for Year 1 and Year 2. Statistical equivalence of these scores is confirmed using the method of Lewis and Lewis,[47] which relies on performing two one-tailed t tests to establish an "equivalency range".[48] The equivalence of these two groups is further bolstered by their overall course scores. The Chemistry-I course scores for these populations were 72.4% ± 10.4% and 72.5% ± 11.2%, for Year 1 and Year 2, respectively. Thus, we have good evidence that the two Chemistry-II experimental populations are equivalent in ability.

The primary measurement tools in this study are the Chemistry-II midterm and final exams. For this purpose, these exams should be as good as or better than other typical examinations. Table 1 presents a summary of test psychometrics for these exams when administered as traditional MC questions via Scantron in Year 1 and when administered in the AUC format via the IF-AT in Year 2. Classical item analysis reveals that indeed these are good exams.[49] An item is said to be *discriminating* when it distinguishes between high-performing and low-performing students. A common means of measuring the discrimination of an item is via correlation between that test item's score and the total test score. For tests with dichotomous items, this is done via the point-biserial correlation coefficient which is equivalent to the Pearson product moment, r. Likewise, r is the appropriate measure for polytomous scores. In general, acceptable exam items should discriminate above r = 0.2, with "good" items exceeding 0.4.[50−52] As shown in Table 1, the Year 1 Scantron exams yield mean item discriminations of 0.40 ± 0.12 and 0.36 ± 0.12. When scored polytomously, the Year 2 IF-AT exams display mean item discrimination of 0.46 ± 0.08 and 0.38 ± 0.11. These values are well above the average for typical classroom exams. For example, DiBattista and Kurzawa report a mean discrimination coefficient (via the point biserial correlation) of 0.27 ± 0.04 for a diverse set of traditional MC tests in a comparable proximate institution.[50] Having items that discriminate well is key to creating a reliable assessment tool.

Test *reliability* is a measure of how consistently the scores reflect what they are attempting to measure.[49,52,53] Thus, high test reliabilities are important here to establish that the exams in this study are good tools for measuring the effects of test format on test score. Traditionally, test reliability is estimated by Cronbach's α, a measure of internal consistency, and these are also displayed in Table 1.[49,54] Test reliability depends on mean item discrimination and scales with the number of items, with "longer" tests being typically more reliable. Thus, to aid in

Table 1. Summary Measures for the Chemistry-II Exams under Typical MC (Scantron) and under AUC (IF-AT) Conditions

Year and Exam Format	Number of Students, N_s	Number of Questions, n_Q	Difficulty, p M ± SD [max, min]	Discrimination Coefficient, M ± SD [max, min]	Reliability, α
Year 1 midterm (Scantron)	353	25	0.57 ± 0.19 [0.84, 0.24]	0.40 ± 0.12 [0.59, 0.11]	0.78
Year 2 midterm (IF-AT)	414	25			
Polytomous			0.64 ± 0.18 [0.86, 0.43]	0.46 ± 0.08 [0.68, 0.28]	0.84
Dichotomous			0.55 ± 0.21 [0.82, 0.27]	0.45 ± 0.08 [0.58, 0.29]	0.83
Year 1 final (Scantron)	346	47	0.67 ± 0.16 [0.94, 0.27]	0.36 ± 0.12 [0.57, 0.03]	0.86
Year 2 final (IF-AT)	406	47			
Polytomous			0.73 ± 0.14 [0.95, 0.37]	0.38 ± 0.11 [0.56, 0.09]	0.87
Dichotomous			0.64 ± 0.17[0.92, 0.26]	0.38 ± 0.11 [0.56, 0.08]	0.87

DOI: 10.1021/acs.jchemed.6b00028
J. Chem. Educ. XXXX, XXX, XXX−XXX

comparing reliabilities between tests of differing lengths, one can use the Spearman−Brown prophecy formula to compute the predicted "adjusted" reliability of any given test, scaled to that of a test comprising equivalently performing items of some standard length.[49] It is becoming more common to compare reliabilities of tests adjusted to 50 items.[50] All of the exams used in this study yield $\alpha \geq 0.78$, with a mean reliability of $\bar{\alpha} = 0.84$. The scaled reliabilities are all above 0.87, with a mean adjusted reliability of $\bar{\alpha}_{50} = 0.89$. Again, this compares favorably both with general guidelines for test reliability that set 0.70 as a benchmark for "reliable" classroom exams,[55] as well as with DiBattista and Kurzawa's reported adjusted reliabilities of $\bar{\alpha}_{50} = 0.74 \pm 0.07$.[50] Thus, the exams used in this study represent good measurement tools for the determination of the effects of test format on test scores.

In the case of the midterms, the mean exam score rose from 0.57 ± 0.19 to 0.64 ± 0.18 in going from Scantron to IF-AT, and in the case of the finals it rose from 0.67 ± 0.16 to 0.73 ± 0.14. This change of 6−7 percentage points is statistically significant (midterm: $t(765) = 5.7$; $p < 0.05$|final: $t(750) = 5.5$; $p < 0.05$) and is a notable result from this study. Simplistically, we could view this change as representing the mean level of student partial knowledge, but it also includes the effects due to test format. These scores include partial credit with the [1, 0.5, 0.1, 0, 0] scheme, as mentioned above. *Post hoc* rescoring these exams with [1, 0, 0, 0, 0] removes the partial credit and provides a measure of a *post hoc* scored dichotomous IF-AT test. We find a statistically significant difference in scores between the Scantron-administered final exam (0.67 ± 0.16) and the identical (*post hoc* dichotomized) IF-AT final exam [(0.64 ± 0.17); $t(750) = 2.3$, $p < 0.05$]. The corresponding difference in the midterm scores is consistent but not statistically significant. The finding that these scores are ∼3 percentage points lower than the Scantron scores is consistent with the notion that *post hoc* dichotomization of IF-AT scores overestimates the effects of partial credit on test score. Prior studies report a difference of 8−13 percentage points when scoring IF-AT with a scheme similar or equivalent to ours versus *post hoc* dichotomized scoring.[35,41−43] Our study indicates that the observed increase in polytomous IF-AT scores over dichotomous Scantron scores comprises two components: an increase of ∼9−10% due to partial credit alone, and a decrease of ∼3% due to behavioral changes associated with the test format ("confidence in guessing", "scratch fever", feedback anxiety, etc.).

Widespread adoption of AUC formats of MC testing that grant partial credit via repeated selection might be inhibited by instructors' fears of grade inflation. Certainly, the availability of partial credit has the potential to significantly increase exam scores such that students who would "fail" a traditional (dichotomous) exam might pass when partial credit is accounted for. In our courses, an exam score of 0.50 is the threshold for failure. In the Year 2 Chemistry-II final exam, 61 students whose dichotomous score would have been below 0.50 passed the exam. This cohort, representing 15% of test takers, saw their exam scores rise by 12.5 ± 2.7 percentage points due to partial credit. Another 27 students failed the exam even with partial credit.

Clearly, scores can increase significantly when partial credit is made available. Nonetheless, as discussed below, ideally this increase should not reflect "grade inflation", which is characterized by a uniform score increase brought about by indiscriminant partial credit. Rather, we have strong evidence that the partial credit granted in our tests is discriminating, and that, as found by Attali, "the difference between initial and revised scores lies in more precise trait measurement and not in measurement of a different trait".[56] Because partial credit is meant to reflect students' states of partial knowledge, it is desirable for the allocated partial credit to strongly correlate with a real measure of knowledge. In a reliable and valid test, the test score should represent the knowledge of the test taker. Thus, a good way to correlate partial credit to knowledge is to compare the rates at which students obtain partial credit with their rates of obtaining full credit; the higher the student's dichotomous score, presumably the more knowledgeable they are. A plot of students' scores with and without partial credit, as shown recently by Grunert et al.,[7] can indeed indicate that the resultant polytomous scoring is within acceptable bounds, but because partial credit overall only comprises a small portion of the total score, it is possible to obtain large, but spurious, correlations between these measures. As a case in point, the correlations between our dichotomous and polytomous test scores in the Year 2 Chemistry-II midterm and final exams are 0.98 and 0.99, respectively. Moreover, a simple correlation between student dichotomous scores and the *total* amount of partial credit they earned is problematic in that the dichotomous score caps the available partial credit; the two are anticorrelated.[7] Nonetheless, the more knowledgeable the student is, the more frequently they should be able to obtain partial credit in a subsequent attempt when they have incorrectly answered a first attempt. Figure 3 presents the correlation between the rate at which students obtain available partial credit and the rate at which they correctly identify the keyed response on their first selection, for both the midterm and final exams. Because there is a discrete set of dichotomous score values, the partial-credit data has been binned by these scores, with error bars representing the standard error in the

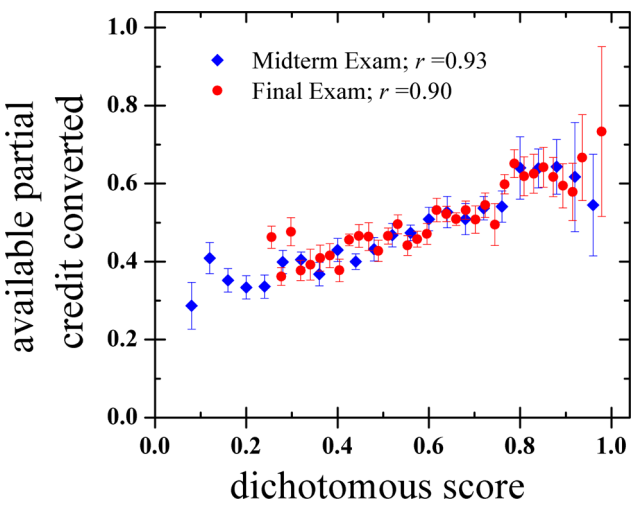


Figure 3. Strong correlation exists between primary student knowledge and their ability to obtain available partial credit. For each of the midterm exam (blue symbols) and final exam (red symbols), a scatter plot of the proportion of *available* partial credit obtained by each student and their exam score when only considering initial responses (dichotomous score) reveals strong correlations of $r = 0.93$ and 0.90, respectively.[57] Dichotomous scores are discrete and have been binned, with error bars representing the standard error in the mean for each bin. Only dichotomous scores with more than one entry are included.

DOI: 10.1021/acs.jchemed.6b00028
J. Chem. Educ. XXXX, XXX, XXX−XXX

mean of each bin. Thus, we only consider score bins with two or more samples in our analysis. Additionally, any students who scored 100% (one or two in each exam) had no opportunity to obtain partial credit and were excluded from the analysis presented in Figure 3.[57]

Two important conclusions can be drawn from the data in Figure 3. First, there is a clear monotonic relationship between "full knowledge" and "partial knowledge", with strong students attaining partial credit at a greater rate than weaker students. (Note the intercept of this graph will not pass through zero because all MC questions, whether on first selection or on subsequent ones, allow some possibility of marks to be assigned due to chance alone.) Second, the data for these midterm and final exams overlap, suggesting that the same cohort approached the two different IF-AT tests in a similar manner when it comes to partial credit. Were there an abundance of random guessing in second or third selections, the relationship seen in Figure 3 would be much weaker, or perhaps even completely flat.

Another interesting piece of evidence that suggests that random guessing in IF-AT tests is not a significant component of the student performance comes from a comparison of 4-option and 5-option items in the Chemistry-II final exams. In Year 1 (Scantron), the 4-option items proved more difficult than the 5-option items. Likewise, in Year 2 (IF-AT), the *post hoc* dichotomized scores are higher for the 5-option items. These differences are 3−4 percentage points, but not statistically significant. The opportunities for guessing on repeat selection in the Year 2 IF-AT exam would imply that the mean polytomous item score for the 4-option items should increase more than it does for the 5-option items.[37] Instead, we find that the mean score of the 4-option items in the Year 2 IF-AT iteration is 0.70 ± 0.17, while it is 0.74 ± 0.13 for the 5-option items. The comparison of these scores is inconsistent with what would be expected from entirely random guessing.[37] Thus, we see no evidence that students are blindly guessing on the IF-AT tests, neither in their initial responses, nor in repeated selections.

The amount to which the IF-AT increases test scores over the traditional MC format clearly depends on the chosen scoring scheme. We have found that compared to a *post hoc* dichotomously scored test (given by [1, 0, 0, 0, 0]), the [1, 0.5, 0.1, 0, 0] scheme raises scores by ∼9−10%. Thus, it is instructive to note the extent to which alternative scoring schemes are anticipated to impact test scores and mean item reliabilities. Table 2 lists the *post hoc* calculated mean item difficulties and discrimination coefficients for various relevant scoring schemes for the Year 2 final exam. For example, we find that for a [1, 1/3, 0, 0, 0] scheme, which we denote as "harsh" for giving less credit for second responses and no credit for third responses, the test score still increases by 5%. The mean item discrimination for this scheme is equivalent to that of [1, 0.5, 0.1, 0, 0], and thus the partial credit is equally discriminating. On the other hand, a "generous" scheme of [1, 0.75, 0.5, 0, 0], is expected to raise test scores by 16% over a *post hoc* dichotomous test. Nonetheless, even with this scheme, the mean item discrimination is only slightly (and not significantly) lower, falling from $\bar{r}$ = 0.34 to 0.31. Only when we test an "irrational" scoring scheme such as [0, 0.1, 0.5, 1, 1] that does not reward primary knowledge and rewards misinformation by giving the most credit for later selections do we find a significant drop in mean item discrimination.[58] Thus, there are ample opportunities to devise scoring schemes that rationally reward partial credit in a manner that faithfully represents partial knowledge, even without experiencing significant grade inflation.

Table 2. Effect of Scoring Scheme on Final Score and Mean Discrimination: Year 2, Chemistry-II Final Exam

Scoring Scheme	Notes	Mean Difficulty, p	Mean Discrimination Coefficient, $\bar{r}$
[1, 0.75, 0.5 0, 0]	"Generous"	0.80[b]	0.31
[1, 0.5, 0.1, 0, 0][a]	As given[a]	0.73[a]	0.34[a]
[1, 0.5, 0, 0, 0]	"Two strikes"	0.72	0.33
[1, 0.33, 0, 0, 0]	"Harsh"	0.69	0.34
[1, 0, 0, 0, 0][a]	Dichotomous[a]	0.64[a,b]	0.33[a]
[0, 0.1, 0.5, 1, 1]	"Irrational"	0.18[b]	0.26[b]

[a]These two schemes are the focus of this article. [b]Indicates a statistically significant difference compared to the as-given scheme (two-tailed t-test; $p < 0.05$).

There are several aspects of this study that limit both its scope and the strength of our findings. The main limitation is that we present a measurement from a single course iteration. While our findings were consistent across two exams (midterm, final), it is possible that our measured score change in going from Scantron to IF-AT would change upon repetition. Repetition, however, would not simply mean looking at scores of arbitrary IF-AT tests, because they may be significantly different than traditional MC tests (for example, if they use integrated testlets). Repetition would require matching Scantron and IF-AT exams and cohorts. Our measurement is currently limited to a single course and discipline, namely, introductory chemistry. Our prior experience with introductory physics suggests little difference in how physics and chemistry students approach the IF-AT, but it is possible that other disciplines outside of the physical sciences would approach this differently. It is also likely that students in upper years would respond differently than freshmen do. Thus, whether our findings apply beyond introductory science courses is uncertain. Finally, because of the nature of repeated selections within the answer-until-correct approach, distractor viability is important. Our tests were representative of acceptable-to-good classroom tests, but even in our exams a significant proportion of items had at least one nonfunctioning distractor. If one were to use IF-AT on a test with poor items that contain several nonfunctioning distractors, they may see a more detrimental effect on the granting of partial credit compared to the initial-response score. In this case, it might be expected that the partial credit will prove significantly less discriminating. Thus, as with all research on the operation of MC test items, this work is most relevant to well-constructed tests.

CONCLUSION

To gain a better understanding of the effects of partial credit on test scores in an answer-until-correct multiple-choice test format, we compared both similar and identical Scantron and IF-AT examinations. We find that a [1, 0.5, 0.1, 0, 0] scoring scheme that grants half-credit if students select the correct response on their second selection and one-tenth credit upon their third selection increases test scores by ∼6−7% compared to a dichotomously administered traditional MC test. That this increase is slightly smaller than that obtained when we simply removed the obtained partial credit (*post hoc* dichotomous marking) suggests that students approach the scratch cards differently than they do Scantron cards, perhaps being slightly

DOI: 10.1021/acs.jchemed.6b00028
J. Chem. Educ. XXXX, XXX, XXX−XXX

less mindful in making initial selections. Nonetheless, we find that, with IF-AT, the partial credit is awarded in a discriminating manner, where the likelihood of any given student obtaining available partial credit is highly correlated with their overall likelihood of initially selecting the keyed option. Partial credit is thus closely tied to partial knowledge with the IF-AT, and its availability improves the test's measurement of content knowledge. While it is difficult to separate random guessing from poor performance, our results demonstrate a strong correlation between overall student performance and ability to correctly answer a question after an initial incorrect response, suggesting that many students are not randomly guessing on subsequent attempts. Additionally, we find obtaining partial credit on 4-option items proves as or more difficult than it is on 5-option items, further supporting the notion that any effects of random guessing are negligible in our tests. This finding thus further diminishes the need for adopting or designing negative-scoring schemes, or for expanding the number of options.

Our findings are most directly relevant for instructors who are considering adoption of answer-until-correct multiple-choice formats for a variety of established pedagogical and technical attributes of partial credit and immediate feedback. We have established a preliminary measure of the anticipated test score increases in moving from typical MC testing, and find that this increase, at ∼6−7 percentage points, is both modest and psychometrically justified. On the basis of these findings we anticipate increased adoption of multiple-choice tests that utilize the IF-AT or any other answer-until-correct response format in introductory chemistry course assessments, including online or computerized delivery.

AUTHOR INFORMATION

Corresponding Author

*E-mail: aaronslepkov@trentu.ca.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank David DiBattista of Brock University for fruitful discussions. We thank Matthew Thompson for the Year 1 Chemistry-II exam data. We also thank the reviewers for insightful and helpful suggestions for manuscript improvement.

REFERENCES

(1) Nicol, D. E−assessment by design: using multiple−choice tests to good effect. *Journal of Further and Higher Education* **2007**, *31*, 53−64.

(2) Scott, M.; Stelzer, T.; Gladding, G. Evaluating multiple choice exams in large introductory courses. *Phys. Rev. ST Phys. Educ. Res.* **2006**, *2*, 020102.

(3) Hartman, J. R.; Lin, S. Analysis of Student Performance on Multiple-Choice Questions in General Chemistry. *J. Chem. Educ.* **2011**, *88* (9), 1223−1230.

(4) Haladyna, T. M. *Developing and Validating Multiple-Choice Test Items*, 3rd ed.; Lawrence Erlbaum: Mahwah, NJ, 2004.

(5) Bodner, G. M. Statistical analysis of multiple-choice exams. *J. Chem. Educ.* **1980**, *57* (3), 188−190.

(6) Scantron Corporation, Eagan, MN; http://www.scantron.com/ (accessed Aug 2016).

(7) Grunert, M. L.; Raker, J. L.; Murphy, K. L.; Holme, T. A. Polytomous versus Dichotomous Scoring on Multiple-Choice Examinations: Development of a Rubric for Rating Partial Credit. *J. Chem. Educ.* **2013**, *90* (10), 1310−1315.

(8) Bush, M. Reducing the need for guesswork in multiple-choice tests. *Assessment & Evaluation in Higher Education* **2015**, *40*, 218−231.

(9) Ben-Simon, A.; Budescu, D. V.; Nevo, B. A Comparative Study of Measures of Partial Knowledge in Multiple-choice Tests. *Applied Psychological Measurement* **1997**, *21*, 65−88.

(10) Frary, R. B. Partial-Credit Scoring Methods for Multiple-Choice Tests. *Applied Measurement in Education* **1989**, *2*, 79−96.

(11) Hutchinson, T. P. Some theories of performance in multiple choice tests, and their implications for variants of the task. *British Journal of Mathematical and Statistical Psychology* **1982**, *35*, 71−89.

(12) Hanna, G. S. Incremental Reliability and Validity of Multiple-Choice Tests with an Answer-Until-Correct Procedure. *Journal of Educational Measurement* **1975**, *12*, 175−178.

(13) Berk, R. A. A consumer's guide to multiple-choice item formats that measure complex cognitive outcomes. In *National Evaluation Systems, From Policy to Practice*; Pearson Education, Amherst, MA, 1996; pp 101−127.

(14) Wilcox, B. R.; Pollock, S. J. Coupled multiple-response versus free-response conceptual assessment: An example from upper-division physics. *Physical Review Special Topics - Physics Education Research* **2014**, *10*, 020124-1−020124-11.

(15) Pressey, S. L. Development and appraisal of devices providing immediate automatic scoring of objective tests and concomitant self-instruction. *J. Psychol.* **1950**, *29* (2), 417−447.

(16) Epstein, M. L.; Lazarus, A. D.; Calvano, T. B.; Matthews, K. A.; Hendel, R. A.; Epstein, B. B.; Brosvic, G. M. Immediate Feedback Assessment Technique Promotes Learning and Corrects Inaccurate First Responses. *Psychological Record* **2002**, *52*, 187−201.

(17) de Finetti, B. Methods for discriminating levels of partial knowledge concerning a test item. *British Journal of Mathematical and Statistical Psychology* **1965**, *18*, 87−123.

(18) Gardner-Medwin, A. R. Confidence Assessment in the Teaching of Basic Science. *Association for Learning Technology Journal* **1995**, *3*, 80−85.

(19) Bush, M. A Multiple Choice Test that Rewards Partial Knowledge. *Journal of Further and Higher Education* **2001**, *25*, 157−163.

(20) Coombs, C. H.; Milholland, J. E.; Womer, F. B. The Assessment of Partial Knowledge. *Education and Psychological Measurement* **1956**, *16*, 13−37.

(21) Brown, A. S.; Schilling, H.; Hockensmith, M. L. The negative suggestion effect: Pondering incorrect alternatives may be hazardous to your knowledge. *Journal of Educational Psychology* **1999**, *91*, 756−764.

(22) Roediger, H. L., III; Marsh, E. J. The Positive and Negative Consequences of Multiple-Choice Testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **2005**, *31* (5), 1155−1159.

(23) Brosovic, G. M.; Epstein, M. L.; Cook, M. J.; Dihoff, R. E. Efficacy of Error for the Correction of Initially Incorrect Assumptions and of Feedback for the Affirmation of Correct Responding: Learning in the Classroom. *Psychological Record* **2005**, *55* (3), 401−418.

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

(25) Epstein, M. L.; Epstein, B. B.; Brosvic, G. M. Immediate feedback during academic testing. *Psychological Reports* **2001**, *88*, 889−894.

(26) QuizSlides is a Web-based application that allows creation of interactive tests, including answer-until-correct multiple-choice questions. https://quizslides.com/features#Overview (accessed Aug 2016).

(27) DiBattista, D. The Immediate Feedback Assessment Technique: A Learner-centered Multiple-choice Response Form. *Canadian Journal of Higher Education* **2005**, *35*, 111−131.

(28) Epstein Educational Enterprises, Inc., Cincinnati, OH; http://www.epsteineducation.com/home/ (accessed Aug 2016).

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

DOI: 10.1021/acs.jchemed.6b00028
J. Chem. Educ. XXXX, XXX, XXX−XXX

(30) Attali, Y. Effects of multiple-try feedback and question type during mathematics problem solving on performance in similar problems. *Computers & Education* **2015**, *86*, 260−267.
(31) Slepkov, A. D.; Shiell, R. C. Comparison of integrated testlet and constructed-response question formats. *Phys. Rev. ST Phys. Educ. Res.* **2014**, *10*, 020120-1−020120-15.
(32) DiBattista, D.; Gosse, L.; Sinnige-Egger, J.-A.; Candale, B.; Sargeson, K. Grading Scheme, Test Difficulty, and the Immediate Feedback Assessment Technique. *Journal of Experimental Education* **2009**, *77*, 311−336.
(33) DiBattiata, D.; Mitterer, J. O.; Gosse, L. Acceptance by undergraduates of the Immediate Feedback Assessment Technique for multiple-choice testing. *Teaching in Higher Education* **2004**, *9*, 17−28.
(34) Epstein, M. L.; Brosvic, G. M. Students prefer the immediate feedback assessment technique. *Psychological Reports* **2002**, *90*, 1136−1138.
(35) Slepkov, A. D. Integrated testlets and the immediate feedback assessment technique. *Am. J. Phys.* **2013**, *81*, 782−791.
(36) Shiell, R. C.; Slepkov, A. D. Integrated Testlets: A New Form of Expert-Student Collaborative Testing. *Collected Essays on Teaching and Learning* **2015**, *8*, 201−210.
(37) Campbell, M. L. Multiple-Choice Exams and Guessing: Results from a One-Year Study of General Chemistry Tests Designed To Discourage Guessing. *J. Chem. Educ.* **2015**, *92* (7), 1194−1200.
(38) Bork, A. Letter to the Editor. *Am. J. Phys.* **1984**, *52*, 873.
(39) Ebel, R. L. Blind Guessing on Objective Achievement Tests. *Journal of Educational Measurement.* **1968**, *5*, 321−324.
(40) Espinosa, M. P.; Gardeazabal, J. Optimal correction for guessing in multiple-choice tests. *Journal of Mathematical Psychology* **2010**, *54*, 415−425.
(41) Persky, A. M.; Pollack, G. M. Using Answer-Until-Correct Examinations to Provide Immediate Feedback to Students in a Pharmacokinetics Course. *Am. J. Pharm. Educ.* **2008**, *72*, 83.
(42) Merrel, J. D.; Cirillo, P. F.; Schwartz, P. M.; Webb, J. A. Multiple-Choice Testing Using Immediate Feedback—Assessment Technique (IF AT®) Forms: Second-Chance Guessing vs. Second-Chance Learning? *Higher Education Studies* **2015**, *5*, 50−55.
(43) Slepkov, A. D.; Godfrey, A. T. K. Partial credit in answer-until-correct multiple-choice test formats. *Applied Measurement in Education*, Submitted January **2016**.
(44) DiBattista, D.; Gosse, L. Test Anxiety and the Immediate Feedback Assessment Technique. *Journal of Experimental Education* **2006**, *74* (4), 311−327.
(45) Testbank to accompany Tro, N.; Fridgen, T. D.; Shaw, L. E. *Chemistry: A Molecular Approach*, Canadian ed.; Pearson: Toronto, ON, 2013.
(46) Both 4- and 5-option items were administered on the same 5-option IF-AT card. In case a 4-option question required the "fifth" spot, the question was swapped with an adjacent question. For the final exam, this meant that two adjacent questions out of 47 were swapped, thereby maintaining near-equivalence of question order between Years 1 and 2.
(47) Lewis, S. E.; Lewis, J. E. The Same or Not the Same: Equivalence as an Issue in Educational Research. *J. Chem. Educ.* **2005**, *82* (9), 1408−1414.
(48) To establish the equivalence between the Chemistry-I final exams of Year 1 and Year 2, we first define an equivalency range q = (−2.5%, 2.5%) within which the two scores would be considered equivalent. Note that this range of ± 2.5 percentage points is a more conservative range than that suggested by Lewis and Lewis,[47] which is based on that required for an effect size below d = 0.2. Such a measure establishes an equivalency range of (−3.0%, 3.0%). We then conduct two one-sided t tests at the a = 0.10 level and deduce:

$$t_1 = \frac{(M_1 - M_2) - \theta_1}{S_P\left(\frac{1}{N_1} - \frac{1}{N_2}\right)} = \frac{(57.5 - 57.9) - 2.5}{14.8\left(\frac{1}{314} - \frac{1}{366}\right)} = 1.86$$

$$t_2 = \frac{\theta_2 - (M_1 - M_2)}{S_P\left(\frac{1}{N_1} - \frac{1}{N_2}\right)} = \frac{2.5 - (57.5 - 57.9)}{14.8\left(\frac{1}{314} - \frac{1}{366}\right)} = 2.57$$

Both of these statistics are larger than the threshold $t(a = 0.10) = 1.28$, and thus, we reject both null hypotheses, thereby establishing equivalence of the two test scores within the defined equivalency range of 2.5%. For a detailed explanation of these parameters and details regarding how to use these values to establish equivalence, see Lewis and Lewis.[47]
(49) Allen, M. J.; Yen, W. M. *Introduction to Measurement Theory*; Waveland Press: Long Grove, Il, 2002.
(50) DiBattista, D.; Kurzawa, L. Examination of the Quality of Multiple-choice Items on Classroom Tests. *Canadian Journal for the Scholarship of Teaching and Learning* **2011**, *2*, No. 4, DOI: 10.5206/cjsotl-rcacea.2011.2.4.
(51) Towns, M. H. Guide To Developing High-Quality, Reliable, and Valid Multiple-Choice Assessments. *J. Chem. Educ.* **2014**, *91* (9), 1426−1431.
(52) Ebel, R. L.; Frisbie, D. A. *Essentials of Educational Measurement*, 5th ed.; Prentice-Hall: Englewood Cliffs, NJ, 1991.
(53) Arjoon, J. A.; Xu, X.; Lewis, J. E. Understanding the State of the Art for Measurement in Chemistry Education Research: Examining the Psychometric Evidence. *J. Chem. Educ.* **2013**, *90* (5), 536−545.
(54) Tavakol, M.; Dennick, R. Making sense of Cronbach's alpha. *International Journal of Medical Education* **2011**, *2*, 53−55.
(55) Nunnally, J. C. *Psychometric Theory*, 2nd ed.; McGraw-Hill: New York, NY, 1978.
(56) Attali, Y. Immediate Feedback and Opportunity to Revise Answers: Application of a Graded Response IRT Model. *Applied Psychological Measurement* **2011**, *35*, 472−479.
(57) We provide here an example illustrating how data are analyzed to form Figure 3: Suppose that in the IF-AT-administered midterm student #1 answered 22 of 25 questions correctly with their initial response. Two of the remaining three questions were answered correctly upon second selection, with the third question requiring four selections to obtain the keyed response. Their dichotomous score is thus 22/25 = 0.88. They had the opportunity to obtain 3 × 0.5 = 1.5 partial-credit points, but only obtained 2 × 0.5 + 1 × 0.0 = 1.0 partial-credit points. Thus, they converted 1.0/1.5 = 0.67 of available partial credit. This, their 0.67 partial-credit "value" would be included in the mean and standard error (ordinate) of the bin (abscissa) with all other students who likewise obtained a 0.88 dichotomous score
(58) Upon initial inspection it might seem counterintuitive that an "irrational" scoring scheme yields a positive measure for mean item discrimination. The reason that the mean discrimination is positive has to do with the fact that the scheme is irrational in its pedagogy but self-consistent in how it awards points. Not getting the answer in the first two responses, and thus earning credit in this irrational scheme, is correlated with not knowing the content. The discrimination, a *de facto* correlation coefficient, is thus positive. Recall that test reliability does not imply validity.

DOI: 10.1021/acs.jchemed.6b00028
J. Chem. Educ. XXXX, XXX, XXX−XXX