2013

Polytomous versus Dichotomous Scoring on Multiple-Choice Examinations: Development of a Rubric for Rating Partial Credit

Megan L. Grunert
Western Michigan University

Jeffrey R. Raker
Iowa State University

Kristen L. Murphy
University of Wisconsin - Milwaukee

Thomas Holme
Iowa State University, taholme@iastate.edu

Polytomous versus Dichotomous Scoring on Multiple-Choice Examinations: Development of a Rubric for Rating Partial Credit

Abstract
The concept of assigning partial credit on multiple-choice test items is considered for items from ACS Exams. Because the items on these exams, particularly the quantitative items, use common student errors to define incorrect answers, it is possible to assign partial credits to some of these incorrect responses. To do so, however, it becomes vital that instructors reach general agreement as to the level of partial credit. Using workshops with instructors, ACS Exams has identified reasons why partial credit could be assigned in an exam set of 70 test items. With partial-credit assignments thus established, polytomous scoring is applied and the effect of such scoring on the overall norm-referenced psychometrics is determined and described. While individual students move within the overall norm, the average influence of polytomous scoring as conceived by the workshops does not substantially change the ability to do norm-based comparisons.

Keywords
first-year undergraduate/general, testing/assessment

JOURNAL OF CHEMICAL EDUCATION

Article

pubs.acs.org/jchemeduc

Polytomous versus Dichotomous Scoring on Multiple-Choice Examinations: Development of a Rubric for Rating Partial Credit

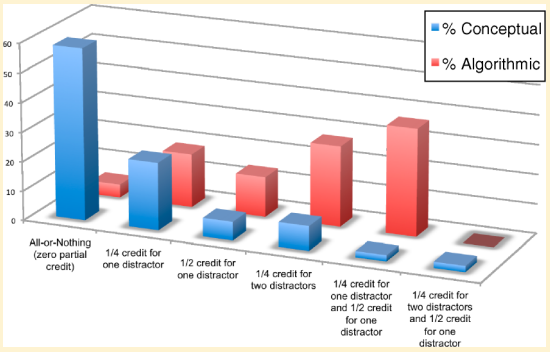
Megan L. Grunert,[†] Jeffrey R. Raker,[‡] Kristen L. Murphy,[§] and Thomas A. Holme*[,‡]

[†]Department of Chemistry, Western Michigan University, Kalamazoo, Michigan 49008 United States

[‡]Department of Chemistry, Iowa State University, Ames, Iowa 50010 United States

[§]Department of Chemistry and Biochemistry, University of Wisconsin−Milwaukee, Milwaukee, Wisconsin 53211 United States

ABSTRACT: The concept of assigning partial credit on multiple-choice test items is considered for items from ACS Exams. Because the items on these exams, particularly the quantitative items, use common student errors to define incorrect answers, it is possible to assign partial credits to some of these incorrect responses. To do so, however, it becomes vital that instructors reach general agreement as to the level of partial credit. Using workshops with instructors, ACS Exams has identified reasons why partial credit could be assigned in an exam set of 70 test items. With partial-credit assignments thus established, polytomous scoring is applied and the effect of such scoring on the overall norm-referenced psychometrics is determined and described. While individual students move within the overall norm, the average influence of polytomous scoring as conceived by the workshops does not substantially change the ability to do norm-based comparisons.



KEYWORDS: *First-Year Undergraduate/General, Testing/Assessment*

INTRODUCTION

Partial credit can be a key feature of free-response examination questions. Many educators use partial credit to describe points awarded to students on examinations or homework for which they have started a problem correctly but ultimately make a mistake or leave out a step that prevents them from arriving at the correct answer. On questions for which students are required to show their work, it is common to either assign points to steps in the problem-solving process or deduct points for missing steps or making mistakes. Using this type of scoring on a multiple-choice examination, such as on a nationally normed American Chemical Society examination, presents several challenges. One such challenge is finding consistency in how and why partial credit is assigned by different instructors.

The American Chemical Society's Examinations Institute (ACS−EI) has been exploring the possibility of partial-credit scoring (i.e., polytomous scoring) as an optional alternative to the current "right or wrong" scoring practices (i.e., dichotomous scoring) on multiple-choice ACS Exams. In realizing the goal of establishing polytomous scoring, a process for the development of a rubric to adjudicate partial credit consistently and reliably was identified and implemented. The first step in the development process was to identify reasons why an examinee would choose an incorrect answer (i.e., the types of errors that lead to a given distractor). The second step involved instructors deciding whether or not they would provide partial credit for these errors if they occurred on a similar, open-response item. To implement this task required a means to delineate types of errors incumbent within distractors to assist instructors in the assignment of authentic and reliable polytomous scoring. This process required the exploration of instructor beliefs and grading practices in order to devise a consensus document that would function despite inherent variability in how instructors view the idea of partial credit. Ultimately, after several iterations of a putative listing of error types, in conjunction with inter-rater reliability studies using chemical education professionals, a valid and reliable rubric has been crafted. The development of the instrument will be reported in this paper, along with results of norm calculations using the polytomous scoring data for a prototype ACS Exam and comparing these results to norm calculations based on dichotomous scoring of the same exam.

Because the adoption of polytomous scoring of multiple-choice test items represents a departure from customary practice, several points need to be made about the assumptions that go along with multiple-choice testing and the type of scoring selected. For a standard dichotomously scored multiple-choice exam, the assumption is that all wrong answers represent an absence of knowledge or an inability to access needed knowledge.[1,2] It does not distinguish between partial knowledge[3] or no knowledge. The goal for this work was to differentiate between answers that demonstrate partial knowledge and those that represent an apparent lack of knowledge through the assignment of partial credit. This task is certainly easier on problems involving calculations rather than conceptual knowledge, primarily because a reasonable assess-

Published: September 16, 2013

ment of partial knowledge is often possible in a multiple-choice format when the content being tested is quantitative in nature.[4] Nonetheless, working with practitioners and using a workshop process allows the identification of proclivities for assigning partial credits for common misconceptions. Thus typical student mistakes can be understood and used to develop an authentic and reliable method for assigning partial credit to general chemistry items, including some that test conceptual knowledge as well.

One substantial challenge with the project as envisioned was the need to develop a method for assigning partial credit on an already established multiple-choice exam. Other researchers seeking to assess partial knowledge through examinations have often developed alternative test structures to accomplish these aims. Ben-Simon, Budescu, and Nevo wrote,[5]

> *The major weaknesses of MC [multiple-choice] tests include susceptibility to guessing and insensitivity to differences between various levels of knowledge (at the individual item level). When faced with a test question, the examinee is, typically, in one of three (subjective) states: (1) the examinee knows the answer fully and with confidence (full knowledge), (2) the examinee knows only part of the answer or is uncertain of the answer (partial knowledge), or (3) the examinee has no knowledge of the answer (absence of knowledge).*

This hypothesized set of states has set the stage for much of the work in this area. Thus, efforts have been focused primarily on the development of alternative testing structures to better differentiate between these knowledge states compared to traditional multiple-choice exams and to then evaluate the ease of use, reliability, and validity of the test structures. Unfortunately, new forms of multiple-choice items introduce the possibility of construct-driven measurement errors, particularly in nationally normed testing, so this method for assigning partial credit is not practical for ACS Exams.

Three types of standard multiple-choice test alternatives were described by Lord and Novick:[6]

1. New item structure and/or response methods
2. Differential item scoring methods (option weighting)
3. Differential test scoring methods (item weighting)

Changing the item structure and/or response methods includes strategies such as instructing students to select all the incorrect (or correct) answers for a question, rather than the single correct answer as is usually done with multiple-choice exams. Credit is assigned based on how many correct (or incorrect) responses are identified, with full credit being awarded only when the student has selected all correct (or incorrect) answers without selecting any incorrect (or correct) answers.[7,8] Other strategies include a confidence measure along with the answer, where examinees can indicate how confident they are in their response, also called *examinee judgment* methods.[9] Maximum credit is awarded when an examinee selects the correct response with the maximal degree of confidence, as this combination is designated as full knowledge. In this system, full misinformation corresponds to the selection of an incorrect answer with the maximal degree of confidence. While this method has shown promising results with its ability to assess partial knowledge, the structure of the exam is quite different from a standard multiple-choice exam.

For the differential scoring methods, two main strategies have emerged from the literature, both based on *direct response* methods, in which the examinee provides a single response. In differential item weighting, item analysis empirically indicates which test items should be more heavily weighted and which should be less heavily weighted, based on measures of item difficulty, validity, or variance. This strategy is sample dependent and has been shown to have no advantage in tests containing more than 10 items.[10,11] In differential option weighting, weights are assigned to alternative answers. The weights assigned to alternative answers can be empirically based on previous or present test administrations, as seen with differential item weighting. Alternatively, experts can assign a priori weights to alternative responses, also called *logical weighting* by Kansup and Hakstian.[12] An example of logical weighting is given in Figure 1 with an explanation of how the weights were assigned based on incorrect processes. This item is representative of an ACS general chemistry item (adapted from the ACS General Chemistry Study Guide).

What is the correct equilibrium constant for the given reaction?

$P_4(s) + 6\ Cl_2(g) \rightleftharpoons 4\ PCl_3(g)$

A. $K_c = \dfrac{[PCl_3]^2}{[Cl_2]^3}$ B. $K_c = \dfrac{[PCl_3]^4}{[P_4][Cl_2]^6}$

C. $K_c = \dfrac{[PCl_3]^4}{[Cl_2]^6}$ D. $K_c = \dfrac{[Cl_2]^6}{[PCl_3]^4}$

Logical Weighting for Scoring a Student's Response

A.	**Zero credit**	Student incorrectly tried to "simplify" the stoichiometric coefficients used in the equilibrium constant expression
B.	**Half credit**	Student incorrectly included a pure solid in the equilibrium constant expression
C.	**Full credit**	Student answered correctly
D.	**Half credit**	Student incorrectly placed the starting materials in the numerator and products in the denominator

Figure 1. Example of logical weighting of a representative ACS general chemistry item.

Logical weighting is the approach taken to develop methods for partial-credit scoring on the ACS examinations. Several reviews have shown this method to improve internal consistency reliability.[10−15] Nonetheless, Frary claims that these methods are unpopular owing to the time needed to develop a weighting system, complicated scoring methods, and challenges in explaining and justifying scoring procedures to examinees.[15] Kansup and Hakstian claim there is "no evidence supporting the sometimes costly and time-consuming practice of differentially weighting item alternatives according to a priori assessments of degree of correctness."[12] This is in accordance with the recommendation of Echternacht that the test item writer use a predetermined system to assign weights to distractors as the exam is developed.[14] These concerns about

dx.doi.org/10.1021/ed400247d | J. Chem. Educ. 2013, 90, 1310−1315

Table 1. List of "Reasons" for Assigning Partial Credit

Reason for Partial Credit	Example
Use of incorrect conversion factors/mol ratios	Multiplying a measurement in meters by 10^{-9} to convert the measurement to units of angstroms
Failure to use conversion factors/mol ratios	Failing to convert from kilograms to grams when calculating molality
Solving for the wrong variable; solving for the variable incorrectly	Calculating the molarity of Mg^{2+} instead of F^- for a solution of MgF_2
Using an incorrect equation; solving the correct equation incorrectly	Predicting an inverse relationship for variables in the $PV = nRT$ equation when the variables are directly related
Error using tabulated data, models, diagrams, or instrumentation	Predicting significant figures for a measurement that are beyond the capability of the instrument (e.g., 10.223 cm when 10.22 cm is the most accurate measurement)
No thought response; misread the question	Choosing an answer that was randomly chosen by the exam writing committee
Failure to understand nomenclature or vocabulary	Choosing the triple point on a phase diagram when the problem asked for the critical point
Failure to understand the concept(s)	Choosing a substitution reaction when the problem asked for the identification of an oxidation−reduction reaction
Failure to translate between symbolic, microscopic, and macroscopic levels	Failing to translate between a balanced equation and the resultant particulate view of matter when the reaction goes to completion
Common/logical misconceptions or mistakes	Multiplying by 2/1 in a stoichiometry problem instead of multiplying by 1/2
Familiarity/recognition/experience	Choosing "tarnish" as a descriptor of a chemical process because the word is unfamiliar (i.e., not covered in most general chemistry courses)

partial-credit scoring for exams operate under different circumstances than have been established via the ACS−EI. No other academic field has a discipline-based testing organization that can marshal resources for exam development, including the possibility of adjudicating appropriate logical weighting models for polytomous scoring. At the same time, the possibility of building a sustainable effort in this type of work requires the development of a reliable rubric to assist in the assignment of weights to distractors during future development of ACS exams. Establishing such a tool will lessen the time and cost needed for this step, while simultaneously simplifying the scoring methods and clearly conveying the scoring procedures to instructors using the exams and the students taking the exams, thus addressing the main criticisms of this technique. Quite importantly, ACS−EI can implement polytomous scoring in tandem with more traditional dichotomous scoring, so instructor comfort levels with a new scoring system can develop over time with experience.

RUBRIC DEVELOPMENT

The 2002 First-Term ACS General Chemistry Exam was selected as the source of multiple-choice examination item data. This exam is no longer an active exam of the ACS−EI, because several newer versions of the exam exist. It is important to note, however, that inactive ACS Exams retain copyright protection, and items from them still cannot be used in other exam instruments or published in any print or electronic venues. It consists of 70 items that cover topics that are normally taught during the first semester of a two-semester college general chemistry course. During active utilization by chemistry instructors, scoring of the examination was dichotomous, with a student's overall score determined by the number of items he or she answered correctly. Norms used here were established from the dichotomous scoring of a national sample of 1178 student performances, from nine institutions, for which individual item answers were submitted.

Development of the partial-credit rubric began with a group of four chemistry faculty and postdoctoral research associates studying the items to determine how students would arrive at each "distractor", the incorrect responses for the 70 examination items. These reasons spanned from incorrect use of a conversion factor to the misapplication of a chemistry concept. This list of "reasons for choosing distractors" was then used on four different occasions by four different sets of individuals to make partial-credit assignments and to revise the rubric. Partial-credit ratings (i.e., data) were collected at a midwestern research institution, and one regional and two national conferences. In total, 23 raters provided partial-credit data, including chemical education graduate students and postdoctoral research associates, chemistry faculty, and developers of ACS Examinations. Participants represented a spectrum of teaching experience and institutional type. Most, but not all, of the participants used ACS Examinations in their courses or at their institutions. Some participants had experience in test development with the ACS−EI, while others did not. Data collection from this large and diverse group was designed to ensure the rubric reflected a representative view of general chemistry instructors on partial credit. Finally, in focus group settings participants worked individually, while in some workshop settings participants worked collaboratively to assign partial-credit scores.

Rating workshops at the regional and national meetings began with an introduction to the project and rubric. Raters worked with ACS−EI staff members to use the rubric to rate two to three sample exam items. Each rater was then asked to provide a partial-credit score for each "distractor". Partial credit was limited to half-credit, quarter-credit, and zero-credit. All distractors for an examination item could be assigned partial credit, zero-credit, or any combination of partial and zero-credit; in other words, all distractors for an examination item could receive partial credit. Raters at the regional and national conferences were also asked to provide reasoning for each partial-credit assignment. During the course of these workshops, the authors took field notes on the raters' discussions and rationales for making partial-credit assignments. These reasons were used to construct the Partial Credit Rubric; 11 reasons for assigning partial credit on multiple-choice items emerged from these data (Table 1). If a distractor was the result of two or more "reasons", the rater was asked to determine a deduction or "change in score" for each reason involved; this gave the authors a measure of the number of points deducted for each "reason". At the conclusion of each rating workshop, the authors reflected on the use of the rubric with the raters; additional recommendations for improving the rubric were gathered at that time. Note that this rubric is not inherently prescriptive as may be common for such templates. In other words, the ideas mentioned allow raters to identify how

dx.doi.org/10.1021/ed400247d | J. Chem. Educ. 2013, 90, 1310−1315

students arrive at incorrect responses, but it does not prescribe the assignment of partial credit when a particular category of error is apparent.

With 70 items to rate, during any given workshop not all items were rated by all participants. Each examination item distractor had partial credit assigned to it by at least 4 raters and up to 12 raters from the four data collection workshops. Possible partial-credit assignment choices were intentionally limited and included half-credit, quarter-credit, and, importantly, no credit. Of the 210 distractors (3 distractors per question; 70 questions), raters agreed on the exact assignments for 58 distractors (27.62%). If only 1 or 2 raters (depending on the number of raters per distractor) dissented, a partial-credit score assignment for the item distractor was made based on the majority of raters; partial credit for 81 distractors (38.57%) was determined with this method. Ratings alone were used to make a total of 66.19% (n = 139) of the partial-credit assignments.

Assignment of the remaining 71 (33.81%) distractors was made by using the reasons and partial-credit values noted by the raters for the other 139 distractors. Using a constant comparison research technique,[16] the authors adjudicated between distractors for which the raters agreed and distractors for which the raters did not agree. For example, items exist about which the raters stated that their reason for partial credit for a given distractor was based on "use of incorrect conversion factors/mol ratios" but disagreed on the value for partial credit. In such cases, distractors on other similar items for which the raters agreed on the value of partial credit for "use of incorrect conversion factors/more ratios" were used to make the partial-credit assignment. This form of assignment of distractor valuation is the most time-consuming and is only possible because of the relatively large number of participants in the workshops for the rubric development.

It is important to note that even when raters agreed on the valuation of a given reason for partial credit for a distractor in one question, that same reason for partial credit may have a different valuation for a distractor in another exam item. In addition, some raters used several reasons for partial credit as a means to subtract points from the total possible; whereas, other raters used other reasons for partial credit as a means to add back points subtracted (e.g., "common/logical misconceptions or mistakes" and "familiarity/recognition/experience"). In other words, expert raters used the rubric in differing ways for each distractor; however, the ability to make two-thirds of the assignments from general agreement suggests that expert raters do agree on partial credit for any given item. This makes it difficult to assign a specific valuation to each of the reasons for partial credit within the rubric. Therefore, it is important that expert raters be used in making such partial-credit determinations and that a single rater is not relied upon to make complete partial-credit assignments.

The partial-credit scores determined by rater agreement, majority agreement, and author assignment were then used to rescore student performances and calculate putative, polytomous norms using the data from the 1178 student responses originally used to establish item statistics for the dichotomous scoring norms. These results differ slightly from the overall norm for this ACS Exam because not all instructors who turn in scores also submit student answers to the individual items. The next section will examine the impact of partial-credit scoring on examinees' percentile ranking, using calculations with the item-inclusive subset of student data for this exam.

IMPACT OF POLYTOMOUS SCORING ON STUDENT PERCENTILE RANKING

The most common concern mentioned by workshop participants during the development of this project was that the use of polytomous scoring on multiple-choice examinations would lead to an artificial inflation of student scores when compared to dichotomous scoring methods. Because norm-based testing is inherently comparative, however, it is not readily apparent how a net raising of raw scores will affect the norm comparison. Thus, student performance scoring based on the assignments noted above using the partial-credit rubric was carried out, and normed scales for both dichotomous and polytomous score were determined.

Partial-Credit Scoring Patterns

Looking at all 70 items of the exam, six scoring options emerged from the inter-rater studies; the percentages of items in each item are provided in Table 2. These options include all-or-nothing scoring (i.e., dichotomous scoring is maintained because no distractors are assigned partial credit), two options with partial credit for one distractor (that vary in the amount of credit assigned that distractor), two options with partial credit for two distractors (again with variable amounts of partial credit), and one exam item that showed partial credit for all three distractors. Items from this examination were separately categorized by three researchers as conceptual (n = 48) and algorithmic (n = 22) questions (α = 0.901 and interclass correlation average measures = 0.899 for the ratings). Items for which complete agreement was not achieved were assigned by majority ratings. The patterns of scoring options differed significantly for conceptual versus algorithmic items (see Table 2 for relative percentages). Conceptual examination items are more likely to show all-or-nothing scoring (58.3% of conceptual items); whereas calculation-based examination items more commonly show partial-credit scoring (63.7% of algorithmic items had partial credit for at least two distractors).

Table 2. Partial-Credit Scoring Patterns of Conceptual and Algorithmic Examination Items

Partial-Credit Scoring Pattern	Conceptual (n = 48), %	Algorithmic (n = 22), %
All-or-nothing (zero partial credit)	58.3	4.5
1/4 credit for one distractor	22.9	18.2
1/2 credit for one distractor	6.3	13.6
1/4 credit for two distractors	8.3	27.3
1/4 credit for one distractor and 1/2 credit for one distractor	2.1	36.4
1/4 credit for two distractors and 1/2 credit for one distractor	2.1	0.0

Scoring Statistics

To clearly designate half- and quarter-credit points for raters, the polytomous scoring of the exam used a 4-point scale (4 points for correct answers, 2 for half-credit and 1 for quarter credit). Partial-credit scoring resulted in a higher percentage mean score (65.3% or 182.8 of 280 possible points, 70 questions worth a maximum of 4 points) than dichotomous scoring (60.7% or 42.5 of 70 possible points) (see Table 3 for additional scoring statistics). This observation is consistent with results from a study by Bauer, Holzer, Kopp, and Fischer, showing that the average score across six exams scored dichotomously was 4−5 points lower than when they were scored with partial credit.[13]

dx.doi.org/10.1021/ed400247d | J. Chem. Educ. 2013, 90, 1310−1315

Table 3. Polytomous and Dichotomous Scoring Statistics

Statistics	Polytomous Scoring: Scaled Score, Max = 70	Dichotomous Scoring: Raw Score, Max = 70
Mean	45.689	42.478
Standard deviation	10.257	11.208
CI (95%)—upper	46.276	43.117
CI (95%)—lower	45.101	41.839

The change in the mean score, however, does not show a complete picture of how scores are affected by changing to polytomous schemes. Figure 2 depicts a graphical relationship

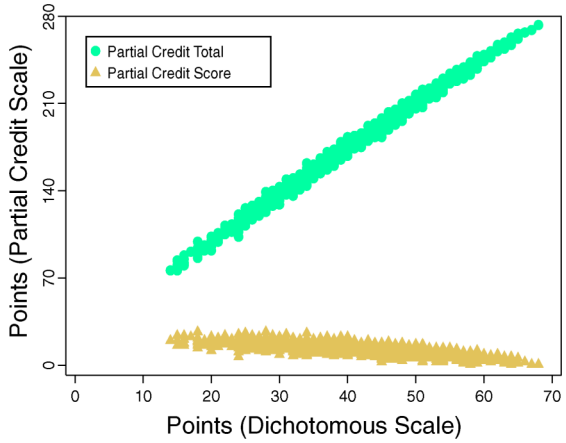


Figure 2. Relationships between polytomous and dichotomous scoring.

between polytomous (*y*-axis) and dichotomous (*x*-axis) scoring. The top set of data (depicted by green circles) shows the overall relationship between the two scoring methods. Considering the change in score from the dichotomous scoring lens, there is variation in polytomous scoring (each dichotomous score has several possible polytomous scores). This variation is tracked visually by the vertical width of the set of green circles; two observations are important here. First, there is no evidence of anomalous and large differences in the two scoring methods. Second, the variability decreases as higher dichotomous scores are achieved. This observation makes sense because at that end of the scale more "correct" responses result in less need to earn partial credit. Figure 2 also includes the amount of partial credit awarded as a function of dichotomous score (i.e., bottom set of data depicted by gold triangles). The pattern of increased partial credit as the dichotomous score decreases was observed, but the data set is rather flat. In other words, the lowest performing students who have the greatest opportunity to gain partial-credit points do not gain much more than the students who have more average performance. The existence of partial credit did not somehow mask the apparent lack of proficiency in the content that is represented by their low scores in the dichotomously scored data.

Normed Scales for Polytomous and Dichotomous Scoring

While the overall picture painted by Figure 2 is that polytomous scoring has a modest effect on measures of student performance, for individual students there can be changes in how they compare within the sample. This is best observed by looking at the percentile ranking using polytomous and dichotomous scoring. Figure 3 depicts this comparison graphically by plotting the polytomous percentile rankings on

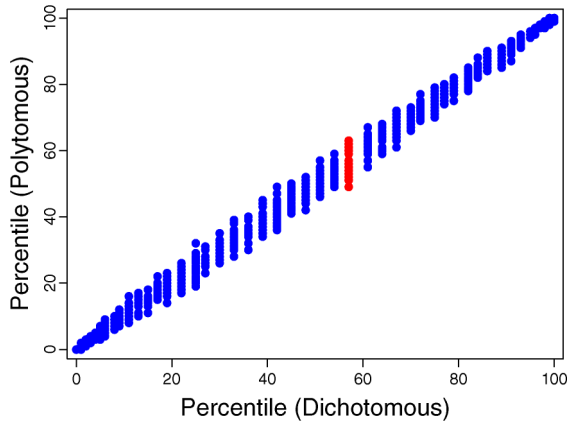


Figure 3. Graphical comparison of polytomous versus dichotomous percentiles. (Data highlighted in red are to assist the reader in viewing how a single dichotomous percentile relates to a range of polytomous percentile rankings.).

the *y*-axis versus the dichotomous percentile rankings on the *x*-axis. Similar to the data for absolute scoring, the data in Figure 3 suggest the largest variation in percentile rankings occurs near the middle of the data (the single largest variance is 13 polytomous percentile values for 1 dichotomous percentile value); a set of polytomous percentiles for a given dichotomous percentile are highlighted in red to assist in interpreting the data. These middle performance students are perhaps the most susceptible to change from the usage of polytomous scoring, because some of them are more often on the right track to answering a question correctly and can select answers for which partial credit will be awarded. Others, who were middle performing because they knew roughly half of the material and were less comfortable with other material, do not demonstrate partial knowledge and gain fewer points via polytomous scoring. These students would fall in this relative ranking method because they would get less partial credit, but importantly, they would likely also drop comparatively in a hand-graded, open-response exam as well. Overall, in comparison, 26.49% of the 1178 examinees' rankings did not change; 35.91% of examinees ranked higher on the polytomous normed scale; 37.6% of examinees ranked lower on the polytomous normed scale. Differences in percentile rankings ranged from positive or negative eight with a normal distribution around zero (i.e., students move up or down zero to eight ranks; mean = 0.026; skew = 0.031; kurtosis = 3.517). This may seem like a large number, but for a 70-item dichotomously scored exam, the change near the middle of the percentile curve for a difference of 1 correct item is often 4−5 percentile points.

CONCLUSION

To best understand the impact of partial-credit, polytomous scoring of multiple-choice exams, it is important to remember that exams from ACS−EI are designed as norm-referenced exams. The exam development process results in items that tend to be relatively good discriminators between low- and high-performing students, for example. Data from this project, therefore, must be considered in light of this comparative lens. In the simplest sense of net score, polytomous scoring on the

dx.doi.org/10.1021/ed400247d | J. Chem. Educ. 2013, 90, 1310−1315

ACS−EI multiple-choice exam results in a higher mean percentage score than dichotomous scoring. Nonetheless, the comparative information, the ability to discern performance levels, is less affected. Low-performing students gain relatively little from the partial credit because they are apparently less likely to make small mistakes that are typically identified as worthy of partial credit via the rubric devised for this project. High-performing students gain little partial credit because they mostly get full credit for their already correct answers. In this sense, the middle-score performers have the potential to benefit more from partial credit than low-score or high-score performers, but clearly not all middle performers do so, as reflected by the variation in percentile rankings. Overall this variation is modest; approximately 26% of percentile rankings are the same with roughly equal proportions increasing and decreasing when switching between the two scoring methods. But the greatest movement lies among the middle-performance-level students. This observation suggests that the addition of partial-credit scoring of multiple-choice items may reflect differences in the style of proficiency of these middle students. Some may have thorough knowledge of only parts of the content domain, while others will have knowledge of more of that domain, though their knowledge is less thorough. This latter group would differentially benefit from partial-credit scoring, and thus in a comparative sense move up in their percentile rank, while the former group would be prone to move down.

Further work needs to be conducted to determine whether this logical-weighting, alternative scoring method is an effective way to evaluate partial knowledge in a multiple-choice format, without the need to change instructions to examinees, item format, or response modes. This study was conducted in a post hoc way with an exam that was not explicitly designed to accentuate the types of different knowledge patterns just hypothesized. Previous work suggests that a priori determinations of distractor scoring by a panel of experts, such as an exam writing committee, are a reliable way of differentiating between full, partial, and absence of knowledge on exam items.[10−15] In addition, interesting investigations could be devised that focus on topics such as (i) what partial-credit scoring could reveal about student's depth and breadth of understanding; (ii) what impact partial-credit scoring has on judgments made about items using classical test theory statistics such as difficulty and discrimination; or (iii) what impact partial-credit scoring has on the percentile rankings for an examination in which significantly more exam items have partial credit available.

A key goal of this project was to determine whether the establishment of a parallel, polytomous-scoring scheme for exams from ACS−EI could be conducted with enough clarity to merit providing this option to users of ACS Exams. The ability to establish an effective rubric that helps identify appropriate partial credit is an important first step. The analysis of existing data sets to see the impact of such a scoring system is the second step. Evidence from this study suggests that a parallel system could be helpful. Thus, the ACS−EI may implement this rubric in scoring general chemistry examinations in the future, providing answer keys and normed percentile rankings for both scoring methods. Moreover, the partial-credit rubric presented here is designed to help the chemical education community in the event that a desire to implement polytomous scoring methods for departmental or course-specific multiple-choice examinations arises. In addition, because the development and reported results were obtained in reference to a general chemistry exam, future research will explore the validity of the partial-credit rubric in scoring other discipline-specific exams (in particular, the organic chemistry exam).

AUTHOR INFORMATION

Corresponding Author

*T. A. Holme: e-mail, taholme@iastate.edu.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We would like to thank Chris Bauer, University of New Hampshire, and Mary Emenike, Rutgers University, for their assistance in developing this project. Additionally, we would like to thank the many individuals who assisted in inter-rater reliability studies and offered exceptional advice for improving the rubric and implementing polytomous scoring for ACS Examinations. Finally, we thank the reviewers for thoughtful suggestions on how the partial-credit rubric might be further utilized as a research tool.

REFERENCES

(1) Bodner, G. M. *J. Chem. Educ.* **1980**, *57*, 188.
(2) Aronson, J. N.; Krause, E. C. *J. Chem. Educ.* **1982**, *59*, 381.
(3) Coombs, C. H.; Milholland, J. E.; Womer, F. B. *Educ. Psychol. Meas.* **1956**, *16*, 13.
(4) Toby, S.; Plano, R. J. *J. Chem. Educ.* **2004**, *81*, 180.
(5) Ben-Simon, A.; Budescu, D. V.; Nevo, B. *Appl. Psychol. Meas.* **1997**, *21*, 65.
(6) Lord, F. M.; Novick, M. R. *Standard theories of mental test scores*; Addison-Wesley: Reading, MA, 1968.
(7) Coombs, C. H. *Educ. Psychol. Meas.* **1953**, *13*, 308.
(8) Dressel, P. L.; Schmidt, J. *Educ. Psychol. Meas.* **1953**, *13*, 574.
(9) Shuford, E. H.; Albert, A.; Massengill, H. E. *Psychometrika* **1966**, *31*, 125.
(10) Stanley, J. C.; Wang, M. D. *Educ. Psychol. Meas.* **1970**, *30*, 21.
(11) Wang, M. W.; Stanley, J. C. *Rev. Educ. Res.* **1970**, *40*, 663.
(12) Kansup, W.; Hakstian, A. R. *J. Educ. Meas.* **1975**, *12*, 219.
(13) Bauer, D.; Holzer, M.; Kopp, V.; Fisher, M. R. *Adv. Health Sci. Educ.* **2011**, *16*, 211.
(14) Echternacht, G. *Educ. Psychol. Meas.* **1976**, *36*, 301.
(15) Frary, R. B. *Appl. Meas. Educ.* **1989**, *2*, 79.
(16) Lincoln, Y. S.; Guba, E. G. *Naturalistic Inquiry*; Sage: Newbury Park, CA, 1985.