Comparison of integrated testlet and constructed-response question formats

Aaron D. Slepkov[*] and Ralph C. Shiell[†]

Trent University, Department of Physics & Astronomy, Peterborough, Ontario K9J 7B8, Canada

(Received 28 March 2014; published 22 September 2014)

Constructed-response (CR) questions are a mainstay of introductory physics textbooks and exams. However, because of the time, cost, and scoring reliability constraints associated with this format, CR questions are being increasingly replaced by multiple-choice (MC) questions in formal exams. The integrated testlet (IT) is a recently developed question structure designed to provide a proxy of the pedagogical advantages of CR questions while procedurally functioning as set of MC questions. ITs utilize an answer-until-correct response format that provides immediate confirmatory or corrective feedback, and they thus allow not only for the granting of partial credit in cases of initially incorrect reasoning, but, furthermore, the ability to build cumulative question structures. Here, we report on a study that directly compares the functionality of ITs and CR questions in introductory physics exams. To do this, CR questions were converted to concept-equivalent ITs, and both sets of questions were deployed in midterm and final exams. We find that both question types provide adequate discrimination between stronger and weaker students, with CR questions discriminating slightly better than the ITs. There is some indication that any difference in discriminatory power may result from the baseline score for guessing that is inherent in MC testing. Meanwhile, an analysis of interrater scoring of the CR questions raises serious concerns about the reliability of the granting of partial credit when this traditional assessment technique is used in a realistic (but nonoptimized) setting. Furthermore, we show evidence that partial credit is granted in a valid manner in the ITs. Thus, together with consideration of the vastly reduced costs of administering IT-based examinations compared to CR-based examinations, our findings indicate that ITs are viable replacements for CR questions in formal examinations where it is desirable both to assess concept integration and to reward partial knowledge, while efficiently scoring examinations.

DOI: 10.1103/PhysRevSTPER.10.020120 PACS numbers: 01.40.Fk, 01.40.gf, 01.50.Kw, 01.40.Ha

I. INTRODUCTION

Constructed-response (CR) questions are a mainstay of introductory physics textbooks and examinations. Often called "problems," these questions require the student to generate an acceptable response by demonstrating their integration of a wide and often complex set of skills and concepts. To score the question, an expert must interpret the response and gauge its level of "correctness." Conversely, in multiple-choice (MC) testing, response options are provided within the question, with the correct answer (the keyed option) listed along with several incorrect answers (the distractors); the student's task is to select the correct answer. Because response interpretation is not required in scoring MC items, scoring is quicker, cheaper, and more reliable [1–3], and these factors contribute to the increasing use of MC questions in introductory physics exams [1,4,5].

With proper construction, MC questions are powerful tools for the assessment of conceptual physics knowledge [4,6], and there are examples of introductory physics final exams that consist entirely of MC questions [1]. These tend to be in universities with large class sizes, where the procedural advantages of MC testing are weighed against any pedagogical disadvantages stemming from an exam that necessarily measures compartmentalized conceptual knowledge and calculation procedures. Conversely, MC questions are not typically used to assess the complex *combination* of cognitive processes needed for solving numerical problems that integrate several concepts and procedures. Such problems involve the integration of a sequential flow of ideas—a physical and mathematical argument of sorts—that can initially seem to resist partitioning into MC items [7,8]. Furthermore, the explicit solution synthesis required by CR questions gives a strong sense of transparency of student thinking that is often lacking in the MC format. For all of these reasons, the use of MC questions for formal assessments in physics education remains limited, and greater exam weight is typically placed on traditional CR questions that involve problem solving and explicit synthesis. Nonetheless, administering MC exams is considerably less time consuming and costly than administering CR exams, and the disparity of cost

[*] aaronslepkov@trentu.ca
[†] ralphshiell@trentu.ca

scales rapidly with the number of students [3]. It is estimated that administering a 3-h integrated testlet (IT) exam employing the Immediate Feedback Assessment Technique (IF-AT) response system as described below costs approximately 0.35/student (including grading and manual data entry), while an equivalent CR exam scored by a single student rater costs at least 7.50/student. Duplicate scoring and/or extensive training of scorers significantly increases these costs. Thus, the cost to administer a CR final exam is on the order of 20 times higher than that of an MC-based exam.

In order to marry the utility of MC with the validity of CR there is a need for new hybrid formats that will provide the procedural advantages of MC testing while maintaining the pedagogical advantages of using CR questions. The recent development of *integrated testlets* (ITs) represents a significant effort to move in this direction [9]. ITs, which are described more fully below, involve the use of MC items within an answer-until-correct format, and are specifically designed to assess the cognitive and task integration involved in solving problems in physics.

A traditional testlet comprises a group of MC items that share a common stimulus (or scenario) and test a particular topic [10–12]. By sharing a common stimulus, the use of testlets reduces reading time and processing as compared to a set of stand-alone questions, and thus improves test reliability and knowledge coverage in a fixed-length examination [10,13]. A reading comprehension testlet provides a classic example of a traditional testlet, with a reading passage being followed by a number of MC questions that probe the student's comprehension of ideas within the passage [12]. A hallmark of traditional testlet theory is the requirement of item independence [10,13], which is necessary to avoid putting students in double jeopardy. That is, because students typically do not receive item-by-item feedback during MC testing, it would be unfair to include an interdependent set of MC questions in the test. Unlike a traditional testlet, an *integrated testlet* is a set of MC items designed to assess concept integration both by using an answer-until-correct framework and by including items with varying levels of interdependence [9]. In an IT, one task may lead to another procedurally, and thus the knowledge of how various concepts are related can be assessed. This approach represents a markedly different way of using testlets. For example, whereas the items in traditional testlets (see, for example, questions 21–24 in Appendix C of Scott, Stelzer, and Gladding [1]) can be presented in any order, the items in an integrated testlet are deliberately presented in a particular sequence.

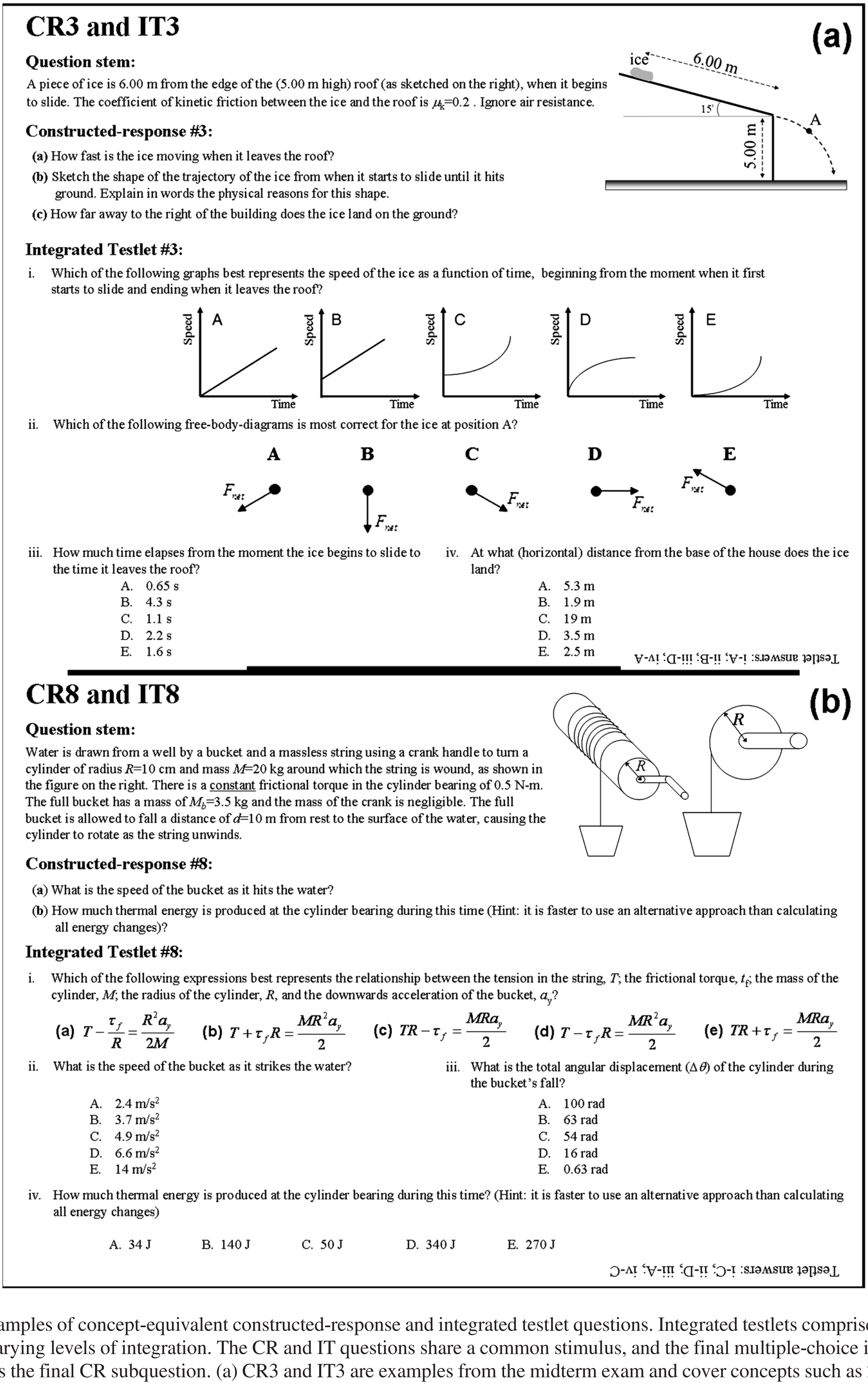
The functional validity of ITs relies on the use of an answer-until-correct response format, wherein the responder is permitted to continue to make selections on a multiple-choice item until the correct response is both identified and can be used in subsequent related items. Certain answer-until-correct response formats, such as the IF-AT [9,14,15], furthermore enable granting of partial credit within MC testing. Thus, we have designed ITs as a close proxy of traditional CR questions: both assess the complex procedural integration of concepts, and both attempt to discern contextual and nuanced knowledge by providing partial credit. Figure 1 presents two examples of traditional constructed-response problems along with two integrated testlets used in the exams described below that use the same stimulus to cover the equivalent conceptual domain.

Despite the converging similarities between CR and IT, some latent differences will remain. For example, physics problems presented in the CR format largely assess concept integration and synthesis, with students implicitly required to generate a tactical plan for solving the problem. In an IT, where several concepts are integrated together to build towards deeper concepts, both the order of the MC items and feedback about the correct answers to individual items suggest to students a possible procedural plan and thus remove some of the synthesis that CR assesses for [compare, for example, CR8(b) and items IT8-iii and IT8-iv in Fig. 1]. To establish how well ITs can act as proxies for CR questions, a direct comparison between the two is needed. The utility of ITs was recently established in a proof-of-principle study that showed that physics exams composed entirely of IF-AT-administered MC items with various levels of integration can be sufficiently valid and reliable for classroom use [9]. Here we report on a head-to-head study in which established CR questions were converted to concept-equivalent ITs, and both CR questions and ITs were simultaneously deployed in midterm and final exams of an introductory "university physics" course. The purpose of this study was to address the following set of related questions: can we adequately convert traditional CR questions to ITs so as to allow for the construction of acceptable IT classroom examinations? How might we go about doing this? How fully is the divide between CR and MC bridged by such an approach, and what is gained and lost when using ITs as a replacement for CR? To address these questions we consider factors such as test statistics, testlet-level psychometrics, CR scoring procedures and interrater reliability issues, anonymous student surveys, and exam deployment costs.

II. METHODS

A. Course structure

A one-term introductory physics course was offered in the fall of 2012 at a primarily undergraduate Canadian university. The course instructor was one of the authors (R. C. S.). The course is a requirement for physics, chemistry, and forensics science majors, and covers topics such as two-dimensional kinematics and mechanics, rotational motion, fluids, and heat. Course delivery followed peer-instruction and interactive-learning principles [16–18], encompassing preclass readings followed by a just-in-time


(a)
ice
6.00 m
15°
5.00 m
A
Speed
Time
A
B
C
D
E
F_{net}
(b)
R

CR3 and IT3

Question stem:

A piece of ice is 6.00 m from the edge of the (5.00 m high) roof (as sketched on the right), when it begins to slide. The coefficient of kinetic friction between the ice and the roof is μ_k=0.2 . Ignore air resistance.

Constructed-response #3:

(a) How fast is the ice moving when it leaves the roof?

(b) Sketch the shape of the trajectory of the ice from when it starts to slide until it hits ground. Explain in words the physical reasons for this shape.

(c) How far away to the right of the building does the ice land on the ground?

Integrated Testlet #3:

i. Which of the following graphs best represents the speed of the ice as a function of time, beginning from the moment when it first starts to slide and ending when it leaves the roof?

ii. Which of the following free-body-diagrams is most correct for the ice at position A?

iii. How much time elapses from the moment the ice begins to slide to the time it leaves the roof?

A. 0.65 s
B. 4.3 s
C. 1.1 s
D. 2.2 s
E. 1.6 s

iv. At what (horizontal) distance from the base of the house does the ice land?

A. 5.3 m
B. 1.9 m
C. 19 m
D. 3.5 m
E. 2.5 m

Testlet answers: i-A; ii-B; iii-D; iv-A

CR8 and IT8

Question stem:

Water is drawn from a well by a bucket and a massless string using a crank handle to turn a cylinder of radius R=10 cm and mass M=20 kg around which the string is wound, as shown in the figure on the right. There is a constant frictional torque in the cylinder bearing of 0.5 N-m. The full bucket has a mass of M_b=3.5 kg and the mass of the crank is negligible. The full bucket is allowed to fall a distance of d=10 m from rest to the surface of the water, causing the cylinder to rotate as the string unwinds.

Constructed-response #8:

(a) What is the speed of the bucket as it hits the water?

(b) How much thermal energy is produced at the cylinder bearing during this time (Hint: it is faster to use an alternative approach than calculating all energy changes)?

Integrated Testlet #8:

i. Which of the following expressions best represents the relationship between the tension in the string, T; the frictional torque, t_f; the mass of the cylinder, M; the radius of the cylinder, R, and the downwards acceleration of the bucket, a_y?

(a) $T-\frac{\tau_f}{R}=\frac{R^2a_y}{2M}$ **(b)** $T+\tau_f R=\frac{MR^2a_y}{2}$ **(c)** $TR-\tau_f=\frac{MRa_y}{2}$ **(d)** $T-\tau_f R=\frac{MR^2a_y}{2}$ **(e)** $TR+\tau_f=\frac{MRa_y}{2}$

ii. What is the speed of the bucket as it strikes the water?

A. 2.4 m/s^2
B. 3.7 m/s^2
C. 4.9 m/s^2
D. 6.6 m/s^2
E. 14 m/s^2

iii. What is the total angular displacement ($\Delta\theta$) of the cylinder during the bucket's fall?

A. 100 rad
B. 63 rad
C. 54 rad
D. 16 rad
E. 0.63 rad

iv. How much thermal energy is produced at the cylinder bearing during this time? (Hint: it is faster to use an alternative approach than calculating all energy changes)

A. 34 J B. 140 J C. 50 J D. 340 J E. 270 J

Testlet answers: i-C; ii-D; iii-A; iv-C

FIG. 1. Examples of concept-equivalent constructed-response and integrated testlet questions. Integrated testlets comprise a set of MC items with varying levels of integration. The CR and IT questions share a common stimulus, and the final multiple-choice item in the IT is the same as the final CR subquestion. (a) CR3 and IT3 are examples from the midterm exam and cover concepts such as 2D projectile motion and kinetic friction. (b) CR8 and IT8 are examples from the final exam and cover concepts such as rotational motion, torque, and work. Note that IT8-iii exists to cue students to the most efficient means of solving IT8-iv. By contrast, such cuing is absent in CR8.

(JIT) online quiz, and in-class clicker-based conceptual tests and peer discussion. Biweekly laboratory sessions were alternated with biweekly recitation sessions at which knowledge of material covered by previous problem sets was tested with 45-min CR quizzes, followed by tutoring of the subsequent problem set. A 2 h midterm exam was administered during week 6 of the 12-week term, and a 3 h final exam was administered shortly after week 12; both exams consisted of a mix of CR questions and ITs. A detailed formula sheet was provided to the students at all quizzes and exams. Exams were collectively worth 50% of a student's final grade. In total, of the 175 initial registrants, 155 students wrote the midterm and 131 students wrote the final exam. Shortly after writing the midterm exam, students were asked to complete an anonymous online survey about their perceptions and engagement with the CR and IT question formats. Of the 155 students who wrote the midterm exam, 105 (68%) completed the survey.

B. Exam construction and scoring

The midterm and final exams were the experimental tools we used to directly compare CR and IT formats. However, because these also needed to be valid and fair evaluation tools within a formal course offering, particular attention was paid to balancing the questions in the experimental design. We designed two sets of complementary midterm exams (blue and red) and final exams (blue and red), where each complementary exam had an equivalent number of CR questions and ITs and covered identical course material, but swapped question formats for each topic covered. Thus, for example, the red midterm comprised, in order, questions CR1, IT2, IT3, and CR4, while the blue midterm comprised the complementary set of IT1, CR2, CR3, and IT4. Each format pair (for example, CR3 and IT3) shared a similar stimulus and covered the same material. The distribution of items among the exams is given in Table I. Examples of complementary questions CR3/IT3 and CR8/IT8 are shown in Fig. 1. Each final exam included six questions—four previously unseen questions and two questions repeated verbatim from the midterm but with altered numerical values. Students were informed in advance that at least one question from their midterm was to be repeated on the final. Students were randomly assigned among the two versions of the midterm and randomized again for the final exams. Thus students were equally divided among the four possible (red and red, red and blue, blue and blue, blue and red) sequence variants.

Author A. D. S., who was not directly involved in teaching the course, designed drafts of the exams and delivered them to the instructor (R. C. S.) two weeks before the scheduled exam time, after which both authors collaborated in editing the exams. Thus, at the time of instruction, the instructor did not know which topics would be tested in the exams. Expanded guidelines employed in constructing concept-equivalent ITs are outlined in the Appendix. In short, all of the major concepts and techniques taught in the course were listed in order of delivery, and a set of constructed-response problems taken from past exams were parsed for overlap of these concepts. Questions were then selected to give the best representation of topic and concept coverage. Rather than designing new CR questions, where possible, we chose to use final exam questions from past years to best assure some construct and content validity. Only one question (CR6/IT6) was newly created for this study because of a gap in topical coverage in recent exams.

TABLE I. Summary of question measures and their placement in midterm and final examinations.

Question		Exam	p'[a]		r'[b]	
			IT	CR	IT	CR
IT1	CR1	midterm	0.63	0.54	0.55	0.64
IT2	CR2	midterm	0.61	0.48	0.69	0.58
IT3	CR3	midterm	0.57	0.43	0.60	0.62
IT4	CR4	midterm	0.54	0.37	0.21	0.69
IT5	CR5	final	0.49	0.41	0.44	0.70
IT6	CR6	final	0.60	0.23	0.51	0.47
IT7	CR7	final	0.63	0.37	0.39	0.53
IT8	CR8	final	0.42	0.30	0.63	0.68
Mean			0.56	0.39	0.50	0.61
(std. dev.)			(0.07)	(0.10)	(0.15)	(0.08)
IT2′	CR2′	final[c]	0.70	0.70	0.44	0.61
IT3′	CR3′	final[c]	0.63	0.61	0.45	0.70

[a] p' is the "item" difficulty parameter and is a measure of the proportion of the available score obtained by the class on a given question.

[b] r' is the *item-excluded discrimination parameter*, it is the correlation between the question score and the total test score exclusive of the question under consideration.

[c] These questions were repeated verbatim on the final exam from the midterm, but with changed numbers. The statistics of these questions are excluded from the combined mean and standard deviation of the other questions

Although, in principle, a testlet can include any number of MC items, each IT used in this study comprised four items, denoted, for example, IT1-i,…,IT1-iv, each with 5 options. Each IT was constructed with its final item being stem equivalent to the final subquestion of its matching CR question.

To enable the answer-until-correct MC response format needed for ITs, we used the commercially available IF-AT [14,15,19], in similar fashion to that described previously [9]. In brief, the IF-AT provides students with immediate confirmatory or corrective feedback on each MC item as they take the test. The IF-AT response sheet consists of rows of bounded boxes, each covered with an opaque waxy coating similar to those on scratch-off lottery tickets. Each row represents the options from one MC question. For each question, there is only one keyed answer, represented by a small black star under the corresponding option box. Students provide their responses by scratching the coating off the box that represents 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 item. Conversely, if no star appears, the student immediately knows that their chosen option is incorrect, and they can then reconsider the question and continue scratching boxes until the star is revealed. It should be noted that the answer key is immutably built into the IF-AT scratch cards and thus the MC questions presented on the exam must be constructed to match the key [20]. This means that the IF-AT is less forgiving of minor errors in test construction than are other MC response techniques. Thus, to aid the proper construction of the tests, the midterm and final examinations were "test driven" by teaching assistants before being administered to the class.

For this comparative study, the exam scoring was designed for simplicity, with all individual MC items worth an equivalent number of marks and each IT worth the same number of marks as each CR question. A major advantage of the IF-AT is that it enables the straightforward use of partial-credit schemes. In our MC items, for reasons outlined in Sec. III D, we gave full credit (5 marks) for a correct response on the first attempt, half-credit (2.5 marks) for a correct response on the second attempt, and one-tenth credit (0.5 marks) for a correct response on the third attempt; no credit was earned for subsequent attempts. In practice, the attempt on which the correct response was attained is inferred from the number of boxes scratched and the presence of a confirmatory star within one such box. For five-option MC items, the marking scheme used can be designated as [1.0, 0.5, 0.1, 0, 0], and the expected mean item score from random guessing is 32%.

To explore the typical reliability of CR scoring, we adopted two commonly found scoring practices; that of utilizing paid student grading and that of instructor grading. CR questions on the midterm exam were scored independently by both authors, who are experienced course instructors. We did not use a common rubric, but we each scored questions in the way we considered most fair and consistent. CR questions on the final exam were scored in duplicate by two paid senior undergraduate students. In this study, they were also given detailed scoring rubrics for two of the six CR questions (CR5 and CR7) and a typical training session explaining how to score fairly and how to use these rubrics. All CR component scores reported herein represent an average of the pair of scores, otherwise known as the *interrater* average score.

C. Exam psychometrics

The more difficult an item, the lower the proportion of available marks that students will earn on it. A widely used item difficulty parameter, p, is traditionally defined as the mean obtained item score. Typically in MC test analysis the scoring is dichotomized and p is simply the proportion of the students that answer the question correctly. In our questions, where partial credit is available, a continuous or polychotomous difficulty parameter p' instead represents the mean obtained question score. p' ranges between 0 and 1, and its value decreases with question difficulty.

At least as important as a question's difficulty is its power to discriminate between more and less knowledgeable students. Whether a question is relatively easy or difficult may be immaterial as long as the item is properly discriminating. Item discrimination is a measure of the correlation between the score on an item and the overall achievement on the test. In the case of dichotomously scored items—such as traditional MC items—the point biserial (PBS) correlation value is traditionally used as a discrimination coefficient [2,21,22]. Here, however, where the availability of partial credit yields polychotomous item scores, the relevant correlation parameter is the Pearson-r.

It should be noted that the correlation between the question scores and the total test scores is not between wholly independent variables [23,24]. Thus, a more pure measure of discrimination is the *item-excluded discrimination parameter*, r', which here is the correlation between the question score and the total test score exclusive of the question under consideration. In all cases, r' is less than r. This distinction becomes less important as the number of questions comprising the total score increases [23], and analysis of standardized tests with ∼100 or more items suffers only marginally by using r rather than r'. Given the number of questions on our exams, r' is certainly the most relevant discrimination parameter for this study. While guidelines exist for interpreting the traditional item discrimination coefficient (PBS) [21,25], there are currently no established guidelines for interpreting the item-excluded discrimination parameter, r'.

III. RESULTS AND DISCUSSION

A. Overview

The mean score on each version of the midterm exam was 52%, and the means were 51% and 50% on the two versions of the final exam. The similarity of mean scores across versions of the exams suggests that the random divisions of the class yielded cohorts with similar overall levels of achievement, and that a comparison of achievement across exam versions is justified.

There is limited data directly comparing achievement in MC and CR formats in the *physics* education literature. While our sample size for the number of responders and the number of questions is somewhat limited, much can be learned from a comparison between how students engaged with concept-equivalent IT and CR questions. Table I lists the p' and r' values for each IT and CR question. It is widely known that students generally obtain lower scores on CR than on MC questions, even when the stems are equivalent [26–28]. This finding is confirmed by our data, where (with the exception of the repeated questions) p' for each IT is larger than that for the corresponding CR question. On average, the IT and CR questions yield mean p' values of 0.56 and 0.39, respectively; the difference being statistically significant, with large effect size [29–31]. The difference in scores between MC and CR items may be attributed to several factors: the added opportunities for guessing available in MC testing; cuing effects resulting from the presence of the correct answer among the MC options; and within our ITs, the fact that feedback provided to students using the IF-AT may enhance performance on subsequent items. Figure 2 presents a scatter plot of p' for each corresponding pair of IT and CR questions. To estimate the plausible increase in p' that arises as the result of random guessing, we can model IT questions as ones where students either know the answer *a priori* or otherwise randomly guess until they find the correct answer (residual guessing). Each question has an "inherent difficulty" assumed to be p'_{CR} and if a student with an innate ability below this difficulty is presented with the question, we assume they resort to random guessing. Thus, while an "equivalency line" would be represented by $p'_{\mathrm{IT}} = p'_{\mathrm{CR}}$, an "equivalency + guessing" line would be represented by $p'_{\mathrm{IT}} = p'_{\mathrm{CR}} + \bar{p}'_{\mathrm{guess}}(1 - p'_{\mathrm{CR}})$, where $\bar{p}'_{\mathrm{guess}}$ is the expected value due to guessing; 0.32 in our case.

As can be seen in Fig. 2, all of the IT questions lie above the equivalency line, but the majority of questions lie between this line and the equivalency + guessing line. The location of any question on this figure is representative of a balance between three possible behaviors: Assuming that the inherent difficulty is indeed p'_{CR}, questions that are predominantly answered via the aforementioned "know it or guess it" approach will be found scattered about the equivalency + guessing line. Questions for which partial knowledge is appropriately rewarded with partial credit will be raised above this line. Questions that contain particularly attractive "trapping" distractors will be lowered below this line because random guessing is interrupted in favor of incorrect responses. The fact that six of eight IT/CR pairs lie between the two lines suggests that distractor trapping (see Appendix for more information) is a significant component of our carefully constructed ITs. The two questions that are found above the top line (representing IT6/CR6 and IT7/CR7) are our best indications of questions that overall show a disproportionate increase in score due to rewarded partial knowledge in the IT format. It should be pointed out, however, that from a probabilistic standpoint alone, of those points found above the top line, more will be found on the low-CR p' side. Overall, this is a simplistic model, but nonetheless may provide a simple means for gauging whether a given IT is more difficult or easier than expected because of its set of distractors.

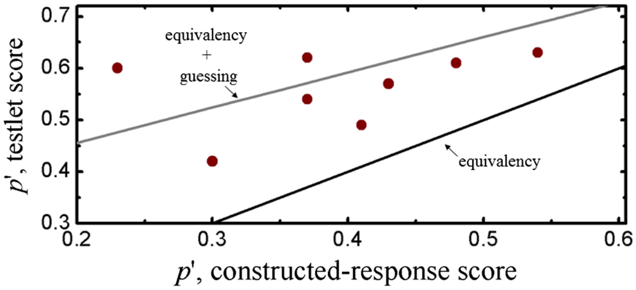


FIG. 2. A comparison of item difficulty parameter p' for each matching pair of constructed response and integrated testlet questions. The solid line labeled equivalency represents the expected relationship between questions of equal difficulty. Similarly, the solid line labeled equivalency + guessing represents the case of equivalent difficulty where those students who do not know the answer choose to guess. Note that the majority of data points lie between these two lines.

There tends to be a moderate-to-strong correlation between students' scores on questions in MC and CR formats [1,32,33]. For example, in a meta-analysis of 56 exams from various disciplines, Rodriguez found a mean correlation coefficient of 0.66 between MC and CR scores [33], while Kruglak found a mean correlation of 0.59 for physics definition questions [32]. Our data are consistent with these findings. Specifically, we found the correlation between students' total IT and CR scores to be 0.70 and 0.83 for the two versions of the midterm and 0.69 and 0.66 for the two versions of the final exam. Thus, there is evidence that, on average, our ITs operate as well or better than traditional stand-alone MC items in approximating CR questions.

A comparison of discriminatory properties of the IT and CR questions (see Table I) also seems to confirm a lesser-known relationship between MC and CR items: When the *scoring* is sufficiently reliable CR questions are typically more discriminating than MC items [27,34]. Our IT and CR questions had mean discrimination scores of 0.50 and 0.61, respectively. With only 8 questions, there is insufficient statistical power to establish statistical significance in discrimination differences between CR and IT [30,31,35]. Nonetheless, the data suggest that our CR questions are more discriminating than ITs, with six out of eight CR questions discriminating at a greater level than their IT counterparts. This may be due in part to guessing that can take place in MC items. Alternatively, as has been identified previously [27,28], because a significant number of students provide blank or irrelevant responses to CR questions, the effective scoring range for CR questions is larger than for ITs. For example, in our study, a score of zero was awarded on an IT only twice out of ≈700 scored ITs, whereas this occurred 21 times on the same number of CR questions. However, a score of 100% was awarded in 9% of all instances of both IT and CR questions. Thus, some loss in discriminatory power may be expected when replacing CR questions with ITs.

While somewhat lower than that for CR, the mean IT item-excluded discrimination parameter r' of 0.50

compares favorably with MC questions typically found on classroom exams. For example, DiBattista and Kurzawa examined 1200 MC items on 16 university tests in a variety of disciplines and found the mean (non-item-excluded) discrimination coefficient to be only 0.25 [36]. Although we feel that the integrated testlet (as opposed to an individual MC item) is the most appropriate unit of measurement, testlet-level psychometric analysis is relatively uncommon [11,37], and, in fact, combining multiple MC items into a testlet increases the discrimination parameter of the testlet over that of the average of the individual MC items comprising the testlet [38]. Thus, for comparison purposes only, we report our mean non-item-excluded discrimination coefficient for the entire set of 40 individual MC items in this study to be an impressive $\bar{r} = 0.45$ [21,39]. Subsequent analysis of our items discounts a link between item interdependence (through item position within testlets) and item discrimination. Thus, the high discrimination values may arise simply due to the care with which all items were written.

B. Test length, composition, and reliability

Test reliability is a measure of how consistently and error-free a test measures a particular construct [22]. A set of test scores that contain a large amount of random measurement error suggests that a repeat administration of the test, or the administration of a suitable equivalent test, could lead to different outcomes. A commonly used measure of test reliability is Cronbach's α, which provides a measure of internal consistency [22,40]. The theoretical value of alpha ranges from zero to one, with higher values indicating better test reliability. For low-stakes classroom exams an $\alpha > 0.70$ is desirable [21,25]. Both versions of our final exam yielded $\alpha = 0.79$, indicating that a mixed-format final exam with three CR and three IT questions can provide satisfactory reliability. In principle, we could assess individually how the IT and the CR portions differentially contribute to the exam reliability, and thus discern which question format is more reliable. However, there is insufficient statistical power to make such a determination in the current study. As a guiding principle, the reliability of a test scales with the number of items (via, for example, the Spearman-Brown prediction formula [41]). Thus, one minor advantage of using ITs over CR may be in the ability to include more items in a fixed-length exam, as preliminary indications suggest that students spend less time completing an IT than they do completing its complementary CR question [42]. The anonymous student surveys support this notion. When asked: "Regardless of which type of question (the testlets or short-answer questions) you found more difficult, which did you spend the most time on?", 48% of students indicated that CR questions took longer to complete, 22% indicated they spent the same amount of time on CR and IT questions, and 27% indicated that ITs took longer to complete. Because test reliability scales with the number of questions, then in order to create optimally reliable IT-only exams, one could add more questions while maintaining test duration. This does not, however, assure that an exam solely comprising ITs will be more reliable than a mixed-format exam containing both IT and CR components. Furthermore, only sufficient gains in question completion times would motivate such an approach. In the past, analysis of which question format, CR or MC, is inherently more reliable has largely proven inconclusive. While some studies find that MC is more reliable, others find the converse [3,33]. We suspect that the relative reliability between MC and CR depends strongly on the balance between the strength of MC item writing and on the consistency of CR scoring [3].

C. Interrater reliability

The manner by which each question contributes to test reliability hinges on any randomness in the scoring of each component. In multiple-choice testing, the contribution of guessing to the total score dominates the discussion of the (un)reliability of the method. On the other hand, while a constructed response may be a more faithful reflection of the responder's state of knowledge, the *interpretation* of that response can be highly subjective, thus diminishing the reliability of the question score. This subjectivity in scoring is inherent to the CR format, but is rarely mentioned when comparing the attributes of CR and MC formats in classroom examinations. As part of our formal comparison of the relationship between CR and IT formats we have assessed the effects of interrater reliability on CR score reliability. As mentioned above, the CR components of the midterm exam were scored in duplicate by two professors, without a shared rubric, while the CR components of the final exams were scored in duplicate by two paid student graders, who shared a formal rubric for two of their six assigned questions. Interrater scoring data are presented in Table II.

Traditionally, a correlation coefficient between the scores of two raters is used as a measure of interrater reliability [28,32,43]. The correlation coefficient for the interrater scoring of every CR question in our study ranges from $r = 0.79$ to 0.95. On first inspection, it may seem that these high correlations imply strong interrater reliability. In the only other similar comparison between MC and CR components on a physics classroom exam we have found, Kruglak reports a similar range of correlations between course instructors scoring in duplicate [32]. The strong interrater correlation does imply that in general raters *rank* students' question scores consistently. However, a closer inspection of student scores suggests a larger amount of both systematic and random variability between raters, as shown in Table II. We find, for example, that the mean difference in question scores ranges from -6 to $+15$ percentage points. Such a sizable mean difference is an indication of interrater *bias*; wherein one rater systematically scores an item higher than the other rater does.

TABLE II. Interrater (IR)[a] reliability for constructed-response questions.

Question		**IR correlation, r**	**IR mean difference**	**IR difference standard deviation**[c]	**IR difference extrema pos./neg.**
MIDTERM[d]	CR1	0.92	-6%[b]	13%	+55%/-20%
	CR2	0.95	+3%[b]	11%	+33%/-25%
	CR3	0.91	+6%[b]	11%	+35%/-25%
	CR4	0.87	-1%	17%	+48%/-33%
FINAL[e]	CR2′[f]	0.92	+11%[b]	15%	+45%/-13%
	CR3′[f]	0.87	+3%	14%	+55%/-38%
	CR5[g]	0.90	+15%[b]	16%	+65%/-5%
	CR6	0.79	+15%[b]	17%	+73%/-13%
	CR7[g]	0.88	+9%[b]	11%	+65%/-30%
	CR8	0.88	+1%	17%	+33%/-45%

[a] All IR differences refer (arbitrarily but consistently) to “scorer 1”-“scorer 2”.
[b] A statistically-significant measure of inter-rater bias, as determined by a paired-item t-test ($p < 0.05$).
[c] A measure of average random error in item scoring
[d] Midterm exam scorers are professors
[e] Final exam scorers were hired senior undergraduates
[f] Repeats on the final exam of midterm CR2 and CR3
[g] Items for which scorers were provided with a detailed scoring rubric

Whereas Kruglak found bias on the order of 4 to 8 percentage points for questions related to physics definitions, we find bias between raters as high as 15 percentage points for paid student scorers for these more traditional physics “problems.” Such bias does not affect the interrater correlation measures, as they are systematic and may not affect the ranking of total scores. On the other hand, the standard deviation in the differences between the scores makes more apparent the measure of random error in scoring. This value ranges from 11 to 17 percentage points. This clearly represents a latent “unreliability” in CR scoring that is not often addressed in the literature. This effect is not merely tied to the common (if nonoptimal) practice of using nonexpert scorers. When the final exam scorers were instructed to use a detailed rubric, the interrater reliability did not improve. Thus, while there is an element of unreliability to what responses mean in MC items, there is also an element of unreliability in the interpretations of responses in CR questions. Classroom tests rarely address this limitation of CR testing. Conversely, scoring “high-stakes tests” often requires great efforts and cost to minimize interrater variability [44], which in large part has motivated the shift towards pure MC testing in standardized tests [3].

D. Validity of partial credit in IT scoring

A key difference between CR and MC testing has traditionally been the means of question scoring. While MC questions are almost invariably scored dichotomously, assessment of partial credit has been a mainstay of traditional physics CR questions. The unavailability of partial credit as a means of rewarding substantial (if incomplete) knowledge is largely seen as a major drawback of traditional MC testing, as it severely limits the assessment of complex knowledge integration. The answer-until-correct framework of IT usage allows for the granting of partial credit for questions in which students initially respond incorrectly, but ultimately respond correctly on subsequent attempts. The validity of such an approach depends on whether the partial credit is being assessed in a discriminating manner; whether it reliably represents some measure of partial knowledge. A previous study of IF-AT-administered exams that utilized a heterogeneous mix of stand-alone items and integrated testlets found that such discrimination is possible [9]. In this current study, as in the former, there is an inverse correlation between the amount of partial credit granted to any given student and their exam score. This is mostly due to opportunity; the top scorers are more likely to get full credit on any question and thus have fewer opportunities to earn partial credit. Nonetheless, the granting of partial credit proves discriminating. To demonstrate this, we consider the likelihood that a student earns the *available* partial credit. Only in cases when a first response is incorrect does a student have the opportunity to earn partial credit. When partial credit is used in a discriminating manner, we expect top students to earn a higher proportion of their available partial credit as compared to the students at the bottom. As a good means of

TABLE III. Analysis of partial credit granted within ITs.

Exam	No. students	Available partial credit converted by top/bottom half of class	t test, p value
Midterm-B	82	50%/35%	0.001
Midterm-R	73	64%/44%	0.001
Final-B	63	56%/47%	0.12
Final-R	68	52%/41%	0.025

TABLE IV. Effects on the average testlet score and discrimination of various hypothetical MC item scoring schemes. All six testlets used on the red and blue final exams are considered. "(change)" denotes deviation from that obtained using the actual, as given, scoring scheme.

Scheme	$\bar{p'}$ (change)	$\bar{r'}$ (change)	notes
[1,0.5,0.1,0,0]	0.58 (N/A)	0.48 (N/A)	as-given
[1,0.5,0,0,0]	0.56 (-0.02)	0.48 (0.00)	"Two-strikes"
[1,0,0,0,0]	0.45 (-0.13)	0.47 (-0.01)	Dichotomous
[1,0.6,0.2,0,0]	0.59 (+0.01)	0.47 (-0.01)	
[1,0.6,0,0,0]	0.59 (+0.01)	0.47 (-0.01)	
[1,0.4,0.2,0,0]	0.57 (-0.01)	0.48 (0.00)	
[1,0.7,0.3,0,0]	0.65 (+0.07)	0.45 (-0.03)	"Generous"
[1,0.3,0,0,0]	0.52 (-0.06)	0.48 (-0.00)	"Harsh"

measuring the discrimination afforded by partial credit, we would ideally use the correlation between the total IT score for each student, scored dichotomously, and the percentage of available partial credit converted. However, the strong inverse relationship between the dichotomously scored total and the opportunity for earning partial credit means that many of the top students do not have much opportunity to earn partial credit, thus making such analysis less robust. Instead, we use a median-split analysis [45,46], which relies on comparisons between the (dichotomously scored) top and bottom 50th percentile groups. As shown in Table III, in all exams the top students converted a higher percentage of available partial credit, as compared to the bottom group. A t test confirms that for three of four exams this difference is significant at the $p < 0.05$ level. We also note that with [1, 0.5, 0.1, 0, 0] scoring, blind guessing for partial credit is expected to yield a conversion rate of 30% of available partial credit. As a cohort, the top scorers obtain partial credit at a much higher rate than this, converting on average 56% of available partial credit, and thus are not likely randomly guessing, but instead are demonstrating partial (or corrected) knowledge of the answers. Furthermore, in all exams, the bottom half of the class also earned partial credit at a higher rate than expected by random guessing.

The choice of marking scheme, and therefore the proportion of partial credit granted, will influence the overall mean test score and may influence the discriminatory power of the granting of partial credit [47]. In this study we used a [1, 0.5, 0.1, 0, 0] scoring scheme, but could have used any number of alternate schemes. There are currently no well-established or research-derived guidelines for the best scoring schemes in examinations that utilize the IF-AT [47]. Over time, we have converged to the current scheme in an attempt to balance a desire to keep the expectation value for guessing sufficiently low as to make passing of the test statistically unlikely with guessing alone, with a desire to prolong students' intellectual engagement with the questions via partial credit incentives. While these two considerations are largely distinct, the choice of scoring scheme must balance the two. From consultation with students we have discerned that offering students an opportunity to "pass" the question by giving them 50% or more on a second attempt is a significant incentive to remain engaged with a question beyond an initial incorrect response. We chose to offer precisely 50% to take advantage of this effect while moderating the overall test score. We then offer 10% on a subsequent correct response in an attempt to keep more students engaged further in the question, again without substantially increasing the overall test score. We have not utilized a scheme that rewards students past a third incorrect attempt because we suspect either that at this point students are likely to revert to random guessing or that whatever partial knowledge they use to answer the question at this point will no longer be discriminating. Whether these considerations are strictly justified is an avenue for future research [47]. Regardless of *a priori* motivations for using this scheme, a *post hoc* analysis of alternate hypothetical scoring schemes proves illuminating. Table IV lists various plausible IF-AT scoring schemes, and compares their effects on the average obtained testlet scores ($\bar{p'}$) and discrimination measures ($\bar{r'}$) for the testlets deployed in the final exams. None of the alternate schemes under consideration include any partial credit beyond the third response because in the actual exam students did not have this option, and thus their fourth or fifth responses are indiscernible. As presented in the table, we considered schemes that range from dichotomous (all or nothing), through a "harsh" scheme that grants a modicum of partial credit for a correct second response, to a "generous" scheme that grants more partial credit for second and third correct responses than in our as-given [1, 0.5, 0.1, 0, 0] scheme. The choice of partial-credit scheme directly affects the average test score. For our exams, the difference between the dichotomous scoring and most generous scoring schemes is 20 percentage points, with an MC component score of 45% for the former and 65% for the latter scheme. As given, the MC component score is 58%. All other schemes considered are much closer to the as-given scheme than to either of the extreme cases. On the other hand, inspection of the effect of the scoring scheme on the discriminatory power of the questions reveals this measure to be quite robust, and $\bar{r}'$ only ranges from 0.45 to 0.48. It is noteworthy that the most generous scoring scheme shows the lowest discriminatory power, while our standard scheme proves as or more discriminating than the others. It should also be noted that the fact that the as-given test proves as discriminating as the (*post hoc*) dichotomously scored test is evidence that partial credit is *at least* as discriminating as the first-response credit. Overall, it appears that all of the plausible schemes considered are viable from a discriminatory standpoint, and thus the main considerations for their adoption are the targeted exam score and student reception.

E. Correlational evidence of similarity between the operation of CR and IT formats

The comparison of how and what CR and MC formats measure in test takers has been an active area of research [1,27,33,43,48,49]. While there is a strong sense from some, including physicists, that the two formats measure fundamentally different things, much research has concluded that there is little evidence to support this notion [3,33]. One of the main research questions addressed by this study is to gauge whether, by the use of ITs, MC can be made more like CR from a content and cognitive domain standpoint. Thus, it is imperative to get a sense for whether IT and CR questions act in fundamentally distinct ways, or whether they are largely acting similarly but with slightly different performance measures. To address this issue we construct a correlation matrix that describes the correlation between each item score on a given exam to every other item on that exam.

Table V presents such a matrix for the two final exams. Overall, all items correlate positively with all other items on the exam, consistent with our discrimination analysis that identified that all CR and IT questions had positive discriminations. The correlations range from 0.14 to 0.62. Comparing the median CR-CR, CR-IT, and IT-IT correlation is highly suggestive that IT and CR items do not behave in fundamentally separate ways. For the red exam, the median item correlations are 0.41, 0.39, and 0.29, respectively, for CR-CR, CR-IT, and IT-IT. Likewise, for the blue exam the values are 0.51, 0.41, and 0.15. Thus, it is clear that while (on average) a CR question behaves most similarly to other CR questions, so too does the average IT. Scores on ITs correlate more closely to those on CR questions than they do to other ITs. This suggests that while the CR format is perhaps measuring what we care about better than does the IT format, the two formats do not measure fundamentally different things. Were these two formats behaving in fundamentally different ways—i.e., accessing different testing “factors”—we would expect the IT-IT median correlations to be higher than the IT-CR median correlations. Factor analysis would be a more direct and robust way to gauge this, but it would also require a much larger study. Thus, while CR and IT questions do not perform to the same level of discrimination, they do not seem to perform distinct measurement tasks, and are hence similar in how they measure the desired construct.

TABLE V. Correlation table for scores of CR and IT questions in the final exams. The upper triangle lists the intratest question correlation coefficients for the red final exam, and the lower triangle lists those for the blue final exam. For example, for the upper triangle the row and column labels 2nd IT each refer to IT3, and for the bottom triangle these refer to IT4, as these are the second IT within each respective exam.

	1st CR	2nd CR	3rd CR	1st IT	2nd IT	3rd IT
1st CR	1	0.62	0.41	0.39	0.17	0.48
2nd CR	0.47	1	0.39	0.51	0.29	0.54
3rd CR	0.56	0.51	1	0.15	0.35	0.40
1st IT	0.50	0.37	0.41	1	0.23	0.29
2nd IT	0.42	0.27	0.39	0.14	1	0.46
3rd IT	0.50	0.28	0.51	0.15	0.40	1

F. Limitations of the study and future directions

This study answers a number of key questions concerning whether or not IT structures can replace traditional CR questions on formal exams. Our study involved ≈150 students, which is triple the size of the previous pilot study [9] and presents for the first time a direct comparison of concept-equivalent CR and IT questions. Nonetheless, many of our results are only suggestive of the differences and similarities between IT and CR. Additional head-to-head testing between CR questions and concept-equivalent ITs is needed to better establish statistical significance between their discriminatory powers. This need is independent of the number of students in the study, and can only be met by deploying and analyzing more CR and IT pairs.

A key difference between CR and IT questions has so far been left unexamined: The procedural cuing implicit in the question order within an IT reduces the testing of solution synthesis that is such a powerful aspect of CR. We have not investigated this nuanced question, which will be addressed in future work. Likewise, the formative assessment nature of ITs has only been hinted at as a key attribute of the tool, [47,50,51] and establishing the extent to which ITs can prove formative will also be addressed in the future. This study aims to compare head-to-head CR and IT formats in an effort to bridge the divide between CR and MC tests. However, no attempt has been made here to compare IT and stand-alone MC questions. This is largely due to our presumption that due to the limited cognitive complexity assessed by typical MC tests, they do not have the *construct validity* we are looking for in a CR physics test. MC tests may *reliably* test something that we are only partially interested in testing. With this study we indicate that a concept-equivalent IT test can measure something much closer to what we want, but possibly with reduced reliability. There has been an ongoing desire to better establish the relationship between CR and MC testing formats by direct comparisons of stem-equivalent questions [33]. Such comparisons, where the only differences between a CR and MC question lies in the availability of response options within the MC item, are the most direct means of measuring differences due purely to the question format, rather than to content or contextual differences. While some of our items are stem equivalent with CR subquestions [for example, IT8-ii/CR8(a) and IT8-iv/CR8(b), as shown in Fig. 1], we cannot directly use our data for a valid stem-equivalent comparison for several reasons: First, not all of our CR subquestions have a strictly stem-equivalent MC item match (for example, IT3/CR3, as shown in Fig. 1). Second, even when subquestions are stem equivalent with a testlet item, there are contextual differences between the items that make such comparisons difficult. For example, the lack of immediate feedback typically leads to subquestions within a CR question that appear more difficult because of aforementioned multiple jeopardy issues. A complete comparison of stem-equivalent CR and IT questions would at the least require either a means for providing immediate feedback in the CR portion of the exams or the introduction of “dummy values” in the CR subquestion stems, in addition to the strict construction of all items as verbatim stem equivalent. This study, on the other hand, compares *concept-equivalent* questions; where the same concept and procedure domains are tested. Thus, this comparison is meant as a more valid comparison between question format than one would get by comparing an arbitrary set of IT and CR questions, but nonetheless presents an incomplete picture of effects of the question format on its discrimination, and the test reliability.

IV. SUMMARY AND CONCLUSIONS

There is a dearth of formal comparisons between multiple-choice and constructed-response question formats in science education. The recent development of integrated testlets—a group of interdependent MC items that share a stem and which are administered with an answer-until-correct response protocol—has been described as a possible replacement for CR format questions in large classroom assessments [9]. In this study, we directly compare the administration of concept-equivalent CR and IT questions in formal classroom exams. We find that scores on ITs are higher than those of equivalent CR questions, but the difference is small and generally within the range accounted for by some of the opportunities for guessing inherent to multiple-choice formats. We find that both CR and IT questions can be highly discriminating and reliable in their assessment of introductory physics knowledge, with the CR format appearing marginally better at both of these measures. A 3-hour mixed-format exam proves to be more than sufficiently reliable for a classroom exam. While a pure CR exam may prove marginally more reliable than a pure IT exam with the same number of questions, because ITs take less time to complete, more questions may be employed to increase the test reliability. A comparison of inter-rater reliability of two individuals scoring CR exams in duplicate reveals that while the score correlations between them is high, there is large latent random and systematic variability in scores. These kinds of data are rare in the literature, and raise important questions of reliability and validity when using multistep CR questions as primary assessment tools. The answer-until-correct response format used for administering ITs allows for straightforward granting of partial credit within the auspices of a multiple-choice test, and we provide evidence that the granting of partial credit is accomplished in a discriminating manner. The ability to assess partial knowledge with IT structures goes a long way towards bridging the divide between CR and MC formats. Finally, an analysis of the correlation between CR and IT scores dispels notions that ITs and CR questions measure distinctly different constructs, but rather suggests that while CR questions are more reliable than IT questions, both types of questions largely measure the same thing. On average, IT scores correlate more closely to other CR scores than to other IT scores.

Beyond any suggestions that for a given exam duration the CR format may prove both more reliable, discriminating, and is *a priori* of higher construct validity, one important comparison remains; that of cost, which is on the order of 20-fold higher for CR than IT exams. We have shown that ITs approximate CR questions and yield comparable measures of reliability, validity, and discrimination, and thus, in light of the disparity in costs, ITs are a viable proxy for CR questions for formal assessments in large classes.

ACKNOWLEDGMENTS

We thank David DiBattista of Brock University for guidance and advice on the experimental design, and for extensive manuscript editing. We acknowledge Greg Hodgson for assistance in testlet construction and experimental design. We also thank Dr. Bill Atkinson and Dr. Rachel Wortis for useful discussion and encouragement. We thank Angie Best of Trent’s Instructional Development Centre for her support. This research was funded, in part, by a Trent University Learning Innovation Grant. We thank Nelson Education and the Dean of Arts and Sciences (Trent University) for co-funding the publication of this paper.

APPENDIX: GUIDELINES FOR INTEGRATED TESTLET DESIGN

To create concept-equivalent IT and CR pairs we started with a set of CR problems taken from past exams, deconstructed the concepts and procedures needed to solve the problem, weighted the importance and difficulty of each part much as one would when constructing a scoring rubric, and created four multiple-choice items that addressed one or two specific conceptual or numerical steps in the solution. Ultimately, the choice of how many items comprise a testlet and which steps in the solution we wish to include in the testlet is based on time constraints and on a targeted difficulty level.

Figure 3 provides a visual map representing this procedure for CR3/IT3 and CR8/IT8, which are reproduced in Fig. 1. We have identified seven nontrivial "elements" in the solution of CR3, and nine in the solution of CR8. In each solution map, we chose four key elements to include as individual MC items, as indicated in the figure. All CR questions used in this study had at least two subquestions, with the solution to the later ones often depending on previous answers. For example, CR3(a) and CR3(b) are independent, but CR3(c) is weakly dependent on CR3(b) and strongly dependent on CR3(a), as depicted in Fig. 3. In creating a 4-item integrated testlet from this question we deemed that CR3(c) is the intended destination of the problem, and thus include it as the final testlet item, denoted IT3-iv. However, the testing of intermediate steps does not necessarily have to follow that of the CR question, and in IT3 we chose three different intermediate elements to test. In this sequence, we do not expect the question to be particularly difficult, and thus the intermediate steps are dispersed and not strongly integrated. It is expected that when the items are strongly integrated and where the final item depends strongly on a particular preceding step, that including this step as an item makes the question easier. This aspect of the answer-until-correct approach mirrors that of Ding's "conceptual scaffolding" question sequences [7,8], where CR questions that involve the particular integration of multiple disparate concepts are preceded by short conceptual MC items that implicitly cue the students to consider those concepts. Thus, Ding's question sequences also utilize an integrated question formalism but without the implementation of immediate feedback or partial credit. We too rely on items within a given testlet to act as scaffolding for other items in the testlet.

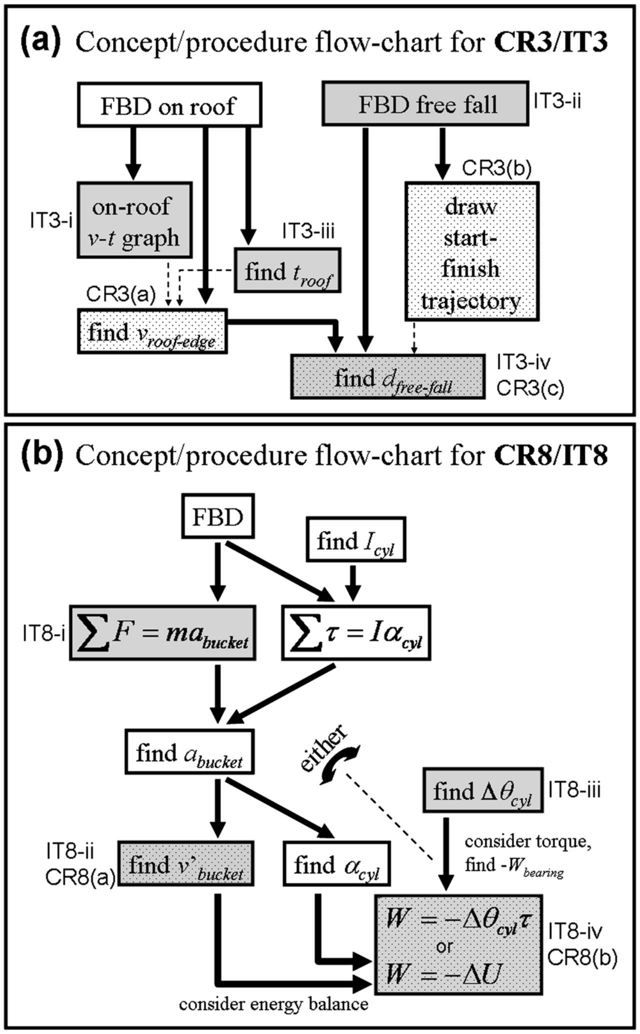

(a) Concept/procedure flow-chart for CR3/IT3
FBD on roof
FBD free fall
IT3-ii
CR3(b)
IT3-i
on-roof v-t graph
IT3-iii
find t_{roof}
draw start-finish trajectory
CR3(a)
find $v_{roof\text{-}edge}$
find $d_{free\text{-}fall}$
IT3-iv
CR3(c)
(b) Concept/procedure flow-chart for CR8/IT8
FBD
find I_{cyl}
IT8-i
$\sum F = ma_{bucket}$
$\sum \tau = I\alpha_{cyl}$
find a_{bucket}
either
find $\Delta\theta_{cyl}$
IT8-iii
IT8-ii
CR8(a)
find v'_{bucket}
find α_{cyl}
consider torque, find $-W_{bearing}$
$W = -\Delta\theta_{cyl}\tau$
or
$W = -\Delta U$
IT8-iv
CR8(b)
consider energy balance

FIG. 3. A conceptual and procedural map of the two concept-equivalent exam structures shown in Fig. 1. The various subitems in the constructed response are labeled and highlighted by dotted stipling. Individual items within a testlet are labeled and highlighted with gray shading. (a) Constructed response question 3(a)–3(c) (CR3) and a 4-item testlet (IT3). FBD = free-body diagram; $d_{\text{free-fall}}$ = horizontal ice landing distance from roof edge. (b) Constructed response questions 8(a),8(b) and 4-item testlet 8 (IT8). cyl = cylinder; v'_{bucket} = bucket's speed at ground height. Arrows indicate which concepts and parameters are needed for developing other concepts and parameters. Two alternate conceptual approaches to the final step are indicated. Unlike in the CR question, the integrated testlet cues and builds scaffolding for an easier approach to the final question.

The issue of how distractors are created is also related to intraquestion scaffolding and discrimination. There are several ways in which distractors can be created: For numerical answers, distractors can be quasirandomly chosen values; they can represent answers obtainable via rational missteps (i.e., identifiable mistakes); and they can be responses that are selected because of their relationship to other distractors. The choice of approach taken for creating any given distractor lies in the assessment objectives of any given question. For example, if a key concept being tested for is the quadratic (as opposed to linear) relationship between two variables, including a distractor that results from a linear analysis may be warranted, as it should aid in discriminating for the key concept. On the other hand, neglecting to trap for such linearity by omitting such a distractor is also tantamount to creating scaffolding within a question. Finally, creating a distractor that results from neglecting to implement a trivial procedure (such as doubling a result) may simply represent a nondiscriminating trap to be avoided. Thus, when choosing discriminators, it is important to also consider the assessment objectives and concept maps underpinning any given question.

The concept of scaffolding is of prime importance to our philosophy of integrated testlets: Ultimately we wish to use MC structures to test how well a student can climb to the apex of a "mountain of knowledge." Typically in a multiple-choice exam, we are relegated to surveying the perimeter of this mountain. With CR questions we often ask the student to climb the mountain, but when a student falters early in the process, they have few tools to assist their climb, and thus we cannot adequately test subsequent progress. With an answer-until-correct integrated testlet we can assess the student's climb from the base to the apex, providing the needed scaffolding as they ascend. Students who do not need the scaffolding get all questions correct on the first try. Some students, however, need help in particular parts of the climb, but can then show that they are able to finish the remainder of the climb without assistance. This is the conceptual framework of the integrated testlet. Consider CR8 [Figs. 1(b) and 3(b)] which deals with rotational dynamics, frictional torque, and work. In the first part of the question, students are asked to solve for the speed of a falling object that is tied to a frictionally coupled rotating cylinder. In the second part, the students are asked about the work done by the friction in the cylinder as the object falls. As shown in Fig. 3(b), there are two conceptually distinct ways to solve CR8(b); the less-efficient method involving the solution to CR8(a). When constructing IT8, IT8-i is a required and important intermediate to CR8(a), with IT8-ii being identical to CR8(a). Then IT8-iii tests a seemingly nonintegrated step that is in fact meant to represent exactly the kind of scaffolding motivated by Ding *et al.* [21]. Finally, IT8-iv is equivalent to CR8(b), thus allowing IT8 and CR8 to test the same conceptual domain. As shown in Table I, CR8 proves to be the second most difficult of all of the exam questions in the course, and IT8 is the most difficult IT given in the course. Thus, the cuing and scaffold building provided by intermediate steps IT8-i and IT8-iii do not significantly simplify the problem, as the IT difficulty value (p') is still below that suggested by Fig. 2. Without direct instructional cuing of how the solution to IT8-iii can help solve IT8-iv, students must still demonstrate that they know how the questions are linked; they must demonstrate the integrated conceptual understanding that is being tested. This notion is further confirmed by the very high value of $r' = 0.63$ for IT8. All eight integrated testlets in our study were created with similar considerations to those outlined above. We considered which steps in the solution to the matching CR question we anticipate will be most difficult and then decided whether to add an intermediate step as an item within the IT. As with IT8, if the solution for a question draws on concepts from different parts of the course we use a mid-testlet question to provide subtle cuing and scaffolding. Because of the aims of the current study we always made sure that the final testlet item was identical to the final CR subquestion. However, because solving an IT may take less time than solving an equivalent CR, and, furthermore, because test takers have confirmatory or corrective feedback at every step, it is certainly possible for an IT to ask questions beyond the scope of the CR, and to do so in a similar time frame on an exam. Thus, ITs could ultimately assess deeper knowledge (i.e., climb a higher mountain) than is viable with a CR question.

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

[2] T. M. Haladyna, *Developing and Validating Multiple-choice Test Items*, 3rd ed. (Lawrence Erlbaum Assoc., Mahwah, NJ, 2004).

[3] H. Wainer and D. Thissen, Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction, Appl. Meas. Educ. **6**, 103 (1993).

[4] G. J. Aubrecht, II and J. D. Aubrecht, Constructing objective tests, Am. J. Phys. **51**, 613 (1983).

[5] S. Tobias and J. B. Raphael, In-class examinations in college-level science: New theory, new practice, J. Sci. Educ. Technol. **5**, 311 (1996).

[6] D. Hestenes, M. Wells, and G. Swackhamer, Force Concept Inventory, Phys. Teach. **30**, 141 (1992).

[7] L. Ding, N. Reay, A. Lee, and L. Bao, Exploring the role of conceptual scaffolding in solving synthesis problems, Phys. Rev. ST Phys. Educ. Res. **7**, 020109 (2011).

[8] L. Ding, N. Reay, A. Lee, and L. Bao, Using conceptual scaffolding to foster effective problem solving, AIP Conf. Proc. **1179**, 129 (2009).

[9] A. D. Slepkov, Integrated testlets and the immediate feedback assessment technique, Am. J. Phys. **81**, 782 (2013).

[10] T. M. Haladyna, Context dependent item sets, Educ. Meas. **11**, 21 (1992).

[11] H. Wainer and C. Lewis, Toward a psychometrics for testlets, J. Educ. Measure. **27**, 1 (1990).

[12] S. G. Sireci, D. Thissen, and H. Wainer, On the reliability of testlet-based tests, J. Educ. Measure. **28**, 237 (1991).

[13] H. Wainer, E. T. Bradlow, and X. Wang, *Testlet Response Theory and Its Applications* (Cambridge University Press, New York, NY, 2007).

[14] M. L. Epstein *et al.*, Immediate feedback assessment technique promotes learning and corrects inaccurate first responses, *Psychol. Rec.* **52**, 187 (2002).

[15] D. DiBattista, The immediate feedback assessment technique: A learning-centered multiple-choice response form, *Can. J. High. Educ.* **35**, 111 (2005).

[16] Eric Mazur, *Peer Instruction: A Users Manual* (Prentice-Hall, Upper Saddle River, NJ, 1997).

[17] C. H. Crouch and E. Mazur, Peer instruction: Ten years of experience and results, Am. J. Phys. **69**, 970 (2001).

[18] D. E. Meltzer and K. Manivannen, Transforming the lecture-hall environment: The fully interactive physics lecture, Am. J. Phys. **70**, 639 (2002).

[19] General information about IF-AT cards is available at http://www.epsteineducation.com/home/.

[20] There is, however, a variety of IF-AT forms with different answer keys available from Epstein Educational—a fact that greatly aids in test security.

[21] L. Ding and R. Beichner, Approaches to data analysis of multiple-choice questions, Phys. Rev. ST Phys. Educ. Res. **5**, 020103 (2009).

[22] R. L. Ebel and D. A. Frisbie, *Essentials of Educational Measurement*, 5th ed. (Prentice Hall, Englewood Cliffs, NJ, 2011).

[23] J. P. Guilford, *Psychometric Methods*, 2nd ed. (McGraw-Hill, New York, 1954).

[24] S. Henrysson, Correction of item-total correlations in item analysis, Psychometrika **28**, 211 (1963).

[25] R. L. Doran, *Basic Measurement and Evaluation of Science Instruction* (Natl. Sci. Teach. Assoc., Washington, D.C., 1980).

[26] W. R. Reed and S. Hickson, More evidence on the use of constructed-response questions in principles of economics classes, *Intl. Rev. Econ. Educ.* **10**, 28 (2011).

[27] M. E. Martinez, A comparison of multiple-choice and constructed figural response items, J. Educ. Measure. **28**, 131 (1991).

[28] D. Barnett-Foster and P. Nagy, Undergraduate student response strategies to test questions of varying format, Higher Educ. **32**, 177 (1996).

[29] The difference in difficulty between matching IT and CR questions is statistically significant. Further, a Wilcoxon signed-rank test shows that there is a large effect size to this difference ($W = 36.0$; $Z = 2.5$; $p = 0.012$, $r = 0.63$). With only 8 pairs of data, such a non-parametric test is most appropriate. Nonetheless, a paired-sample t test also yields $p < 0.05$).

[30] F. Wilcoxon, Individual comparisons by ranking methods, Biometrics Bull. **1**, 80 (1945).

[31] S. Siegel and N. J. Castellan, *Non-Parametric Statistics for the Behavioral Sciences*, 2nd ed. (McGraw-Hill, New York, 1988), pp. 87–90.

[32] H. Kruglak, Experimental study of multiple-choice and essay tests. I, Am. J. Phys. **33**, 1036 (1965).

[33] M. C. Rodriguez, Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations, J. Educ. Measure. **40**, 163 (2003).

[34] R. W. Lissitz, X. Huo, and S. C. Slater, The contribution of constructed response items to large scale assessment: Measuring and understanding their impact, *J. Appl. Test. Tech.* **13**, 1 (2012).

[35] With only 8 pairs of data, there is insufficient statistical power to establish statistical significance in the discrimination parameters of CR questions and their matching ITs. A Wilcoxon signed-rank test yields $W = 7.0$, $Z = 1.54$; $p = 0.12$, $r = -0.385$, and we cannot reject the null hypothesis. Note that this finding is also supported by a paired-sample t test, which yields $p > 0.05$.

[36] D. DiBattista and L. Kurzawa, Examination of the quality of multiple-choice items on classroom tests, *Can. J. Scholar. Teach. Learn.* **2**, 4 (2011).

[37] While Wainer *et al.* (see references, above) have developed a deep theory of testlet psychometrics, they focus largely on item response theory (IRT), and we are not aware of examples in the literature that discuss classic item analysis for testlet-level psychometrics. Our study is insufficiently large to consider analysis with IRT or factor-analytic models.

[38] In every one of the more than 50 testlets we have analyzed over the past few years we have found that the testlet-level r' is higher than the average of the r' values of the comprising MC items. We have not been able to find this discussed in the classic testlet theory literature and are currently studying this further.

[39] Since this measure is being reported for direct comparison with that found in other MC exams, we calculated this value by only including multiple-choice items in our midterms and finals. That is, we removed all CR items and calculated the item-total Pearson-r for the 8 MC items of each midterm and the 12 MC items of each final, then we averaged all 40 values.

[40] N. M. Webb, R. J. Shavelson, and E. H. Haerte, Reliability coefficients and generalizability theory, in *Psychometrics*, edited by C. R. Rao and S. Sinharay (Elsevier, New York, 2006), p. 81–124.

[41] J. P. Guilford, *Psychometric Methods*, 2nd ed. (McGraw-Hill, New York, 1954), p. 354.

[42] The most reliable estimates we have of the time each student spends on a given CR and IT come from follow-up studies of quizzes administered in the recitation period. Here, tutorial leaders administer sequentially a CR and an IT (or vice versa) in a fixed amount of time, and have indicated that on average students hand in the IT portions much earlier than the CR versions.

[43] G. R. Hancock, Cognitive complexity and the comparability of multiple-choice and constructed-response test formats, J. Exp. Educ. **62**, 143 (1994).

[44] Measurement Incorporated, Maryland High School Assessment 2002 Scoring Contractor Report, October 2002, available at http://www.marylandpublicschools.org/NR/rdonlyres/099493D7-805B-4E54-B0B1-3C0C325B76ED/2386/432002ScoringContractorsReport.pdf.

[45] The use of median-split analysis as a replacement for correlational analysis is becoming increasingly controversial. While the practice is nearly never justifiable, a rare exception is one in which the data counts are heavily skewed toward one extreme or another, as is in our case of lack of partial-credit availability to the top exam scorers. See MacCallum *et al.* [46], 2002.

[46] R. C. MacCallum *et al.*, On the practice of dichotomization of quantitative variables, Psychol. Meth. **7**, 19 (2002).

[47] D. DiBattista, L. Gosse, J.-A. Sinnige-Egger, B. Candale, and K. Sargeson, Grading scheme, test difficulty, and the

immediate feedback assessment technique, J. Exp. Educ. **77**, 311 (2009).

[48] S.-Y. Lin and C. Singh, Can multiple-choice questions simulate free-response questions?, AIP Conf. Proc. **1413**, 47 (2012).

[49] R. E. Bennett, D. A. Rock, and M. Wang, Equivalence of free-response and multiple-choice items, J. Educ. Measure. **28**, 77 (1991).

[50] R. E. Dihof *et al.*, Provision of feedback during preparation for academic testing: Learning is enhanced by immediate but not delayed feedback, *Psychol. Rec.* **54**, 207 (2004).

[51] G. M. Brosvic *et al.*, Efficacy of error for the correction of initially incorrect assumptions and of feedback for the affirmation of correct responding: Learning in the classroom, *Psychol. Rec.* **55**, 401 (2005).