2016

Examining the effects of testwiseness in conceptual physics evaluations

Seth Devore
West Virginia University, stdevore@mail.wvu.edu

John Stewart
West Virginia University, jcstewart1@mail.wvu.edu

Gay Stewart
West Virginia University, gay.stewart@mail.wvu.edu

Digital Commons Citation

Devore, Seth; Stewart, John; and Stewart, Gay, "Examining the effects of testwiseness in conceptual physics evaluations" (2016). *Faculty Scholarship*. 917.
https://researchrepository.wvu.edu/faculty_publications/917

Examining the effects of testwiseness in conceptual physics evaluations

Seth DeVore,[*] John Stewart,[†] and Gay Stewart[‡]

Department of Physics and Astronomy, West Virginia University, Morgantown, West Virginia 26506, USA

(Received 13 May 2016; published 10 November 2016)

Testwiseness is defined as the set of cognitive strategies used by a student that is intended to improve his or her score on a test regardless of the test's subject matter. Questions with elements that may be affected by testwiseness are common in physics assessments, even in those which have been extensively validated and widely used as evaluation tools in physics education research. The potential effect of several elements of testwiseness were analyzed for questions in the Force Concept Inventory (FCI) and Conceptual Survey on Electricity and Magnetism that contain distractors that are predicted to be influenced by testwiseness. This analysis was performed using data sets collected between fall 2001 and spring 2014 at one midwestern U.S. university (including over 9500 students) and between Spring 2011 and Spring 2015 at a second eastern U.S. university (including over 2500 students). Student avoidance of "none of the above" or "zero" distractors was statistically significant. The effect of the position of a distractor on its likelihood to be selected was also significant. The effects of several potential positive and negative testwiseness effects on student scores were also examined by developing two modified versions of the FCI designed to include additional elements related to testwiseness; testwiseness produced little effect post-instruction in student performance on the modified instruments.

DOI: 10.1103/PhysRevPhysEducRes.12.020138

I. INTRODUCTION

Multiple-choice tests are a widely used means of evaluation and are very important in physics education research (PER) because of the extensive application of research-validated conceptual instruments. Multiple-choice instruments also suffer numerous potential weaknesses inherent to using multiple-choice items [1–4]. Some of these weaknesses may be exploited by students to improve their chances of selecting the correct answer regardless of the student's content knowledge [5]. These weaknesses include strategies such as examining the length of the available options [6] and converging on the correct answer based on sets of similar options or analysis of other patterns of available options [3,7]. Testwiseness is the collective use of cognitive strategies to exploit weaknesses inherent to the format or characteristics of the test to achieve a higher score [5,8–10]. Testwiseness has been acknowledged in the literature for over sixty years as a potential factor affecting reliability [11].

These weaknesses have led to a considerable number of item writing rules [12,13] and several assessment level rules [14]. These rules have been developed to aid in the construction of multiple-choice items and structuring of multiple-choice assessments that minimize the effect of the application of testwiseness. Despite the existence and dissemination of these rules through textbooks and articles, many items developed for and included in introductory level texts violate one or more of these item writing rules [15]. Unfortunately, very little research confirming the validity and reliability of these item and test writing strategies exists [16]. Haladyna and Downing performed an extensive search for theoretical and empirical studies that supported their taxonomy of 43 multiple-choice item and assessment writing rules and found that for nearly half of the rules no supporting research could be identified [17]. One study suggested that by understanding and exploiting the rules employed in structuring the answer key of the Scholastic Aptitude Test (SAT) students can increase their verbal SAT score by between 10 and 16 points, which is considerably more than the increase in score from participation in formal coaching programs [14,18].

Well-established item writing rules include key balancing, avoidance of related distractors, and avoidance of "none of the above" or "all of the above" distractors. Key balancing refers to the practice of selecting items that produce an instrument with an approximately uniform distribution of answer choices. Item location within the instrument can also be adjusted to avoid consecutive sequences of the same answer choice. Key balancing is important because of students' expectation of a balanced key developed through experience with standardized testing and student answering patterns which may select distractors in certain locations preferentially. Students unsure of the correct answer may have patterns in selecting answers where they preferentially select the middle option

[*] stdevor@mail.wvu.edu
[†] jcstewart1@mail.wvu.edu
[‡] gbstewart@mail.wvu.edu

(option "c" for a five-option item), "central bias," select the first item ("a"), "primacy," or the last item ("e"), "recency." Students can also use strategies to limit the number of options from which they make a random selection by focusing either on the selection containing the most words or by focusing on pairs of selections that are parallel or opposite. Testwiseness has been shown to be important for the overall performance of the evaluation. Many standardized instruments try to produce a balanced set of multiple-choice options with correct answers randomly distributed across the available options, a balanced key [14].

The effect of the inclusion of a none of the above (NOTA) or an all of the above option has been extensively explored [19–21]. All of the above options are rare in the physics education research instruments examined; therefore, this paper will focus on NOTA options. Haladyna and Downing [17] found the use of a NOTA option to be the third most commonly investigated item-writing and/or test-writing rule among the 43 they identified. While some disagreement regarding the effect of NOTA on item validity exists [20], most sources argue that it should be avoided in the development of evaluation tools [13,17].

Another form of testwiseness involves students' bias towards selecting multiple-choice answers based on the option's position [22–25]. Attali and Bar-Hillel measured a predisposition to select the central answers both by instructors when generating multiple-choice questions and by students when guessing the answer to a question [26]. In contrast, other studies have found that primacy and recency appear to have a stronger effect on responses to multiple-choice items resulting in the first and last alternatives being predominately selected [27,28]. Further studies noted differences in the effect of position bias for various types of questions [29].

Another potential element of testwiseness involves commonalities that exist between sets of multiple-choice answers [5,25,30] which can occur when the correct answer is written first and the distractors are then written to share characteristics with the correct answer. This can result in a system of distractors that are similar or opposite to the correct answer [5,25,30].

Testwiseness strategies for improving scores on examinations for items where the student either has no or limited content knowledge are most effective when students understand the item and assessment writing rules discussed above. In most cases, students have no reason to believe that physics assessments conform to these rules; however, most students in today's physics classes have been exposed to years of standardized testing utilizing instruments that conform to good assessment construction practices. During informal discussions, many students reported awareness of features of physics evaluations that activate their testwiseness strategies. Some students report specific instruction into testwiseness strategies as part of their preparation for standardized examinations. Because of this conditioning, it seems possible that, if presented with a multiple-choice instrument that violated some of the common assessment construction rules, a students' pattern of responses might be altered leading to changes in the distribution of item answers or overall evaluation scores.

Testwiseness-influenced changes in overall scores or item scores could have important implications for PER. Research in physics education often involves the use of multiple-choice conceptual instruments and compares pretest and post-test scores using a statistic, the normalized gain, that forms a composite of the two scores. Because testwiseness effects should be more prevalent on the pretest, they may influence the interpretation of the normalized gain. Testwiseness-affected distractors may affect item-level analyses and alter the interpretation of changes from pretest to post-test [1,17,20,26].

Testwiseness plays the role of both a metacognitive strategy students employ for monitoring their performance during an examination and a cognitive strategy which is applied for specific items. Students report awareness of the overall key balancing of examination instruments which takes the form of an awareness of highly unbalanced sets of responses (e.g., many more a's) or long sequences of the same responses. This awareness may raise unnecessary doubts in students that have consequences for the outcome of the examination.

This study complements and extends the existing literature on testwiseness and PER. With potential gains from utilizing testwiseness strategies rivaling the gains from other more traditional strategies, a more extensive examination into the effect of these strategies is warranted. The degree to which items in research-validated physics instruments may be vulnerable to testwiseness effects is investigated through a survey of widely used PER instruments. Testwiseness has been exclusively studied as a general effect common to all disciplines; however, little exploration of discipline-specific testwiseness has been conducted. The existence of testwiseness effects specific to instruments involving scientific reasoning is explored. Finally, the degree to which testwiseness influences the results of some of the most widely deployed PER instruments is examined.

This study addresses the following questions. (i) To what extent do the instruments used in physics education research conform to well-established item writing rules? (ii) Are there testwiseness effects that are specific to scientific assessment instruments? (iii) Does student application of testwiseness affect the outcome of the PER instruments at the item or instrument level?

II. METHODS

A. Quantitative testwiseness effects

Testwiseness research has predominately focused on the performance on broad evaluations requiring little topical knowledge; as such, research into testwiseness effects specific to individual disciplines is rare. We seek to demonstrate the existence of testwiseness effects specific to quantitative disciplines. We will demonstrate that students show an aversion to selecting the distractor zero, "zero bias." An extensive review of the literature pertaining to testwiseness and item writing rules revealed only rules that were tangentially related to zero options (e.g., avoid specific determiners such as always, never, all, none) [15]. Zero distractors will be defined as "0" or zero options, as well as options which imply zero (e.g., "the object will not move" implies zero velocity). Zero bias will be analyzed along with the more extensively studied NOTA bias after an analysis of the extent to which these two effects are present in widely used PER evaluation instruments. NOTA distractors were most commonly identified as options which explicitly said none of the above or none of these (sometimes followed by additional explanation such as "The ball falls back to ground because of its natural tendency to rest on the surface of the earth" or "The elevator goes up because the cable is being shortened, not because an upward force is exerted on the elevator by the cable"). Options which similarly identified none of the other four options as correct were also included as NOTA distractors (such as "other", "not enough information is given to answer the question", or any statement including "cannot determine" sometimes followed by additional text such as "without knowing the forces $-Q$ exerts on the two negative q's"). Also options in which no force is exerted are included as NOTA options such as "none of the forces. Since the chair is at rest there are no forces acting upon it." The zero and NOTA options identified in the FCI and CSEM are shown in Table I.

B. Analysis of research-based assessments

To explore the degree to which item writing rules designed to reduce the effects of testwiseness are applied in the construction of instruments in PER, 12 introductory level physics assessments were examined. Assessments were selected that spanned a variety of introductory physics topics including mechanics, electricity and magnetism, waves, and optics. The assessments include many of the most widely used instruments in PER. Each instrument was analyzed for several factors including the number of NOTA options, the number of zero options, and the distribution of correct answers for each of the assessments that included only five-option questions. Analysis of the distribution of correct answers for the remaining assessments was uninformative because each contained questions with varying numbers of answer choices.

TABLE I. Zero and NOTA options identified in the FCI and CSEM.

	Zero	NOTA
FCI	11 E, 15 E, 16 E, 29 E	13 E, 17 E, 20 E, 29 E
CSEM	1 E, 2 E, 8 A, 9 A, 10 E, 12 C, 13 E, 14 B, 15 E, 19 E, 20 E, 21 E, 24 E, 27 E, 28 E, 31 A	3 E, 4 E, 5 E, 6 E, 8 E, 9 E, 16 E, 23 E, 26 E, 29 E, 30 E, 32 E

Table II shows that all of the 12 assessments examined contain NOTA options with varying frequency. Across the 12 assessments analyzed, NOTA options were available in approximately one-third (33.6%) of all questions. All but one of the assessments examined contained at least one instance of a zero option. Zero options appear in slightly over one-third (34.5%) of all of the questions examined.

The distribution of correct answers for the five assessments that were comprised of only five-option multiple-choice questions shows that many of these assessments appear to have been developed to have an approximately even distribution of correct answers. One notable exception was the distribution of correct answers present in the FCI; option (b) was the correct answer for ten of the questions in the FCI while the average of other options was only five.

These 12 assessments only represent a small sample of those available in physics education research but are representative of some of the most used in PER. Despite the popularity of these assessments, all ignore one or more commonly accepted item writing rules. The degree to which these features affect the student selection of incorrect answers and, therefore, the interpretation of the patterns of student answering is explored in the following sections.

C. Data sets

The analysis which follows utilizes four data sets collected at two universities: U1 and U2. Data set 1 and 2 were both collected at a large midwestern U.S. land-grant university (U1) serving between 15 000 and 25 000 students. Data set 3 and 4 were collected at a large eastern U.S. land-grant university (U2) serving approximately 30 000 students.

Data set 1 was collected from students in introductory calculus-based mechanics and electricity and magnetism classes that were administered the Force Concept Inventory (FCI) [31,43] and Conceptual Survey of Electricity and Magnetism (CSEM) [37], respectively. The data were collected from the Fall 2001 semester to the Spring 2014 semester. Each of these assessments was administered as both a pretest, prior to instruction, and as a post-test, after instruction resulting in over 6000 responses to the FCI (6617 pre and 6241 post), over the course of 26 semesters, and approximately 3000 responses to the CSEM (2992 pre and 3074 post), over the course of 19 semesters. The students received credit for a good-faith effort on the FCI pretests. The FCI post-test and the CSEM pre- and post-tests were graded for credit. For two semesters, Spring 2006 and Fall 2006, the students were asked to report for each question whether they were sure of their answer or were

TABLE II. Comparison of the number of questions with NOTA options and zero options, as well as the distribution of correct answers for all assessments that consist solely of five-option multiple-choice questions.

Assessment name	Assessment subject	Questions with NOTA options	Questions with zero options	Total number of questions	Correct answers A B C D E
FCI [31]	Mechanics	4	4	30	5 10 5 4 6
FMCE [32]	Mechanics	30	43	47	···
EMCS [33]	Mechanics	2	0	25	5 5 5 5 5
RRMC [34]	Mechanics	12	12	30	6 7 8 4 5
ECA [35]	Mechanics	18	14	28	···
MBT [36]	Mechanics	9	4	26	6 5 6 4 5
CSEM [37]	Electricity and Magnetism	12	16	32	6 6 5 7 8
DIRECT [38]	Electricity and Magnetism	2	3	29	···
BEMA [39]	Electricity and Magnetism	6	2	31	···
MCS [40]	Electricity and Magnetism	10	3	30	6 6 6 6 6
CUE-CMR [41]	Electricity and Magnetism	12	6	24	···
MWCS [42]	Waves and Optics	2	15	22	···

guessing. Additional analysis of this experiment was presented in Stewart and Stewart [44].

The FCI is composed of questions which address 6 common Newtonian concepts distributed across 30 questions. It is intended to force students to choose between Newtonian concepts and common sense alternatives [31]. The CSEM is composed of 32 questions which address 10 electricity and magnetism concepts as well as Newton's third law and how it applies to electricity and magnetism problems [37]. Both tools are designed to be administered as both a pretest and post-test to measure conceptual learning gains.

Data set 2 was collected from students in the calculus-based introductory electricity and magnetism class at U1 over the course of 10 semesters from the Fall 2007 semester to Spring 2012. This data set contains all multiple-choice questions given in the class including homework, lecture quizzes, laboratory quizzes, and in-semester examinations. These questions were developed by the teaching staff for use in the course and did not consist of questions from the FCI, CSEM, or any other standardized assessment. These questions include a mixture of qualitative and quantitative items developed for the classes studied. A total of 1851 students were included in this data set with 243 084 total responses to multiple-choice questions recorded. For the purposes of this study, only five-option multiple-choice questions were analyzed. These questions were a mix of qualitative (30%) and quantitative (70%) questions. All questions were given for credit post-instruction.

Data set 3 was collected in the introductory calculus-based electricity and magnetism course at U2. The CSEM was administered as a pretest and post-test from Spring 2011 to Spring 2015 to roughly 2000 students (2278 pre and 1753 post). Students received credit for a good faith effort on both the pretest and post-test.

Data set 4 was collected in the introductory calculus-based mechanics and electricity and magnetism classes at U2 during the Spring 2015 semester. In both classes, each student was administered one of three versions of the FCI modified as a post-test at the end of the class in an attempt to elicit testwiseness effects. The modifications are described in Sec. III. Students received credit for a good faith effort on these examinations. A total of 475 students completed the assessment with approximately 160 students completing each instrument.

III. RESULTS

A. NOTA and zero bias

The analysis of the effects of NOTA and zero options began by examining two assessment instruments designed to evaluate student understanding of Newtonian mechanics (FCI) and electricity and magnetism (CSEM). While there were instances of these testwiseness-affected options being the correct answer, our analysis will focus on the questions for which all testwiseness-affected options were distractors. This will allow comparison of the strength of these testwiseness-affected distractors with the strength of other distractors. Any problems that contained both zero and NOTA distractors were also excluded resulting in the removal of two questions from the CSEM analysis. After these removals, 4 NOTA and 4 zero distractors remained in the FCI and 10 NOTA and 13 zero distractors remained in the CSEM. For five-option multiple-choice questions, such as those used in both the FCI and CSEM, the average likelihood of a student who answers incorrectly to randomly select any of the four available distractors is 25%. Any deviation in the rate of selection of these distractors from that of a uniform distribution, which should occur if all distractors are equal in strength, should be indicative of the distractor's relative strength. The distribution of incorrect student responses for questions in data sets 1 and 3 containing either NOTA or zero distractors is presented in Table III.

TABLE III. Total number and percentage of students selecting the testwiseness-affected distractor and other distractors in questions which have a NOTA distractor or zero distractor in data sets 1 and 3. For "other distractors," the number is the total number of selections for all three other distractors and the percentage is the average percentage for one of the three. Superscripts * denotes $p < 0.05$, ** denotes $p < 0.01$, and *** denotes $p < 0.001$ based on a χ^2 test of difference from a random distribution.

	None of the above			Zero		
	NOTA distractor	Other distractors	χ^2	Zero distractor	Other distractors	χ^2
FCI pretest (U1)	1118 (5.8%)	18 054 (31.4%)	3757.0***	811 (4.5%)	17 119 (31.8%)	4009.6***
FCI post-test (U1)	249 (3.4%)	7102 (32.2%)	1831.3***	278 (4.5%)	5860 (31.8%)	1371.8***
CSEM pretest (U1)	1508 (9.1%)	15 142 (30.3%)	2257.1***	4552 (18.4%)	20 238 (27.2%)	582.5***
CSEM pretest (U2)	1123 (9.0%)	11 347 (30.3%)	1701.4***	3440 (17.8%)	15 842 (27.4%)	527.1***
CSEM post-test (U1)	1154 (14.6%)	6774 (28.5%)	461.2***	3083 (23.0%)	10 339 (25.7%)	29.5***
CSEM post-test (U2)	547 (7.8%)	6800 (30.7%)	1207.5***	1886 (15.4%)	10 398 (28.2%)	609.7***

Students selected the NOTA distractors in the FCI and the CSEM at a statistically significantly lower rate than the other distractors. While the selection of the NOTA distractor was highest for the CSEM post-test, it was still less than the 25% selection rate that is expected from random chance. Zero distractors in the FCI were also selected at a very low rate, less than 5%, for both the pretest and post-test. For the CSEM, these rates were considerably closer to random selection for both the pretest and post-test at both U1 and U2. The CSEM pretest zero distractor selection rates were substantially less than 25% at both institutions. The CSEM post-test selection rate for U2 was also substantially less than 25%, but the post-test rate for U1 was 23%. This result was still statistically significantly different from 25% [$\chi^2(1, N = 13422) = 29.51$, $p < 0.001$] but represents a small effect. The class instructor for the U1 course in which the CSEM was administered reported explicitly confronting zero bias in his lectures, thus potentially affecting the outcomes for the CSEM post-test.

These results support previous work showing that NOTA options are weak distractors. Further, zero options are identified as weak distractors demonstrating the existence of testwiseness effects specific to quantitative disciplines. Student bias against selecting either NOTA or zero options should be strongly considered when developing an assessment. The inclusion of either of these options as a distractor could result in students randomly selecting the correct answer more often than intended. While this cannot be demonstrated for the above analysis, testwiseness may also partially suppress the selection of the correct answer when the correct answer is NOTA or zero.

B. The effect of testwiseness on overall scores

Testwiseness may influence item scores by making it either more likely that a student randomly selects the correct answer when the testwiseness option is incorrect or less likely that the student selects the correct answer when it is testwiseness affected. Using the values in Table III as well as the distribution of questions in the FCI and CSEM with zero and NOTA as either a correct or incorrect answer, an estimate of the cumulative effect of zero and NOTA answers can be calculated for the average student. To calculate the estimate, we make the following assumptions: (i) student avoidance of a testwiseness-affected distractor will increase the probability of selecting each of the other available options equally, which will increase the likelihood of selecting the correct answer, and (ii) student avoidance of a testwiseness-affected option that is correct will produce a similar effect decreasing the likelihood of selecting the correct answer. Applying these assumptions, the net effect on the FCI was an increase in score of 0.58% from NOTA distractors and of 0.55% from zero distractors. The calculated change in the CSEM was 0.20% from NOTA options and 0.05% from zero options. These small net effects for the FCI and CSEM indicate that for these instruments overall scores are not substantially affected by testwiseness; however, the effect on the interpretation of item-level results could be substantial.

C. Misconceptions

An extensive body of research has shown that some students bring strongly held misconceptions to physics classes [45,46] and that these misconceptions are often not removed by instruction [47,48]. The instruments employed by this study were constructed to contain distractors that represented the results of applying common misconceptions. As such, the low rate of selection of zero or NOTA options may result because the zero or NOTA option does not represent a common misconception. To explore this effect, the two semesters of data in data set 1 that asked the students to express whether they were sure of their answer or guessing when answering were analyzed. The pretest and post-test results of the NOTA and zero-affected questions is presented in Table IV.

For the FCI, those students who were guessing on the question selected the zero and NOTA distractors more frequently than students who reported being sure of their answer; however, for the guessing students the rate of selection of the NOTA and zero options was still

TABLE IV. Distractor distribution for students who are “sure” of their answer and students who are “guessing” in data set 1. Total number and percentage of students selecting the testwiseness-affected distractors and other distractors in questions which have a NOTA distractor or zero distractors. For category other distractors the number is the total number of selections for all three other distractors and the percentage is the average percentage for one of the three. Superscripts * denotes $p < 0.05$, ** denotes $p < 0.01$, and *** denotes $p < 0.001$ based on a χ^2 test of difference from a random distribution.

	None of the above			Zero		
	NOTA distractor	Other distractors	χ^2	Zero distractor	Other distractors	χ^2
FCI pretest (Sure)	22 (3.6%)	587 (32.1%)	148.6^{***}	17 (3.2%)	511 (32.3%)	133.6^{***}
FCI pretest (Guess)	40 (8.5%)	428 (30.5%)	67.6^{***}	31 (6.3%)	462 (31.2%)	92.1^{***}
FCI post-test (Sure)	13 (3.6%)	353 (32.1%)	89.8^{***}	12 (4.6%)	249 (31.8%)	57.9^{***}
FCI post-test (Guess)	8 (13.3%)	52 (28.9%)	4.4^{*}	5 (8.5%)	54 (30.5%)	8.6^{**}
CSEM pretest (Sure)	16 (7.1%)	208 (31.0%)	38.1^{***}	84 (23.4%)	275 (25.5%)	0.5
CSEM pretest (Guess)	121 (10.8%)	999 (29.7%)	120.4^{***}	331 (20.2%)	1311 (26.6%)	20.5^{***}
CSEM post-test (Sure)	37 (9.1%)	368 (30.3%)	54.4^{***}	129 (22.1%)	454 (26.0%)	2.6
CSEM post-test (Guess)	28 (12.4%)	197 (29.2%)	18.9^{***}	130 (28.2%)	331 (23.9%)	2.5

substantially less than would be predicted by chance. This is exactly the pattern one would expect if some of the students who were sure of their answers were using a misconception not represented by the NOTA or zero distractor.

For the CSEM, the NOTA results are similar to those found in the FCI with guessing students selecting the NOTA option less frequently than predicted by chance but more frequently than the sure students. The zero option results for the CSEM were less clear. For the pretest, the guessing students select the zero option a statistically significant 5% less often than predicted by chance [$\chi^2(1, N = 1642) = 20.53$, $p < 0.001$]. The sure students’ selection rate of the zero option for the pretest was not significantly different than that predicted by chance. Neither the sure nor the guessing students selected the zero option at a significantly different rate than that predicted by chance on the post-test. The pretest results seem to indicate that the zero option represents a common misconception on some CSEM problems—this would explain the difference in the zero option results between the FCI and the CSEM. The explicit confrontation of the zero option by the instructor may have modified the students’ application of this testwiseness strategy.

D. Position bias

A student may also select multiple-choice answers in situations where the correct answer is unknown based on the position of the answer choice. This effect will be called “position bias.” Position bias can interact with NOTA or zero bias because these options are often placed as the last option. The effect of position-bias was examined in five-option, multiple-choice homework, quiz, and test questions collected in data set 2. Three classes of questions were selected for examination, those with a NOTA-affected option (e), those with a zero-affected option (e), and those with neither testwiseness effect present in any of the options. The first two classes of questions were selected for examination as a result of the prevalence of both NOTA and zero as option (e). NOTA appeared almost exclusively as option (e) with only a few problems containing a NOTA option as one of the other four options. Zero appears considerably more often than NOTA across the other four available options, but still appears as option (e) roughly twice as often as it appears in the sum of the other four options. The selection of the third class of questions, those with neither testwiseness effect, was made to determine students’ distribution of selection in the absence of other testwiseness effects. With the predominance of NOTA and zero appearing as option (e), it was impossible to disentangle the distribution of student selection resulting from position bias from the effects of NOTA and zero bias across all questions. Examining questions without NOTA and zero options provides an opportunity to determine the effects of the position bias alone, as well as the cumulative effects present in the first two classes of questions.

For each of these three classes of questions, all instances of students answering incorrectly as well as the distractor that was selected were recorded. From this set of incorrect responses, instances of questions with correct answers in each of the possible positions [(a), (b), (c), (d), and (e)] were equally, and randomly, sampled to ensure that no bias was introduced because of a prevalence of correct answers in any one position. The distribution of incorrect responses is presented in Table V. If the position of the distractor had no bearing on the likelihood of it being selected, the distribution of incorrect answers should be evenly distributed across the five available options resulting in an average of 20% of the students selecting each option.

Option (e) was selected less often than the other four options for all three classes of questions in Table V. The NOTA distractor continued to be selected at a substantially lower rate than other distractors in the (e) position. The students selected the zero option at approximately the same rate of other options (e) but at a significantly lower rate than would be predicted by chance. Options (b) and (c) were

TABLE V. Total number of students selecting each distractor for five-option multiple-choice questions in data set 2 under three conditions (E was a NOTA option, E was a zero option, or neither NOTA nor zero options were present). For each condition equal numbers of incorrectly answered questions were sampled with options A, B, C, D, and E as the correct answer. All distributions of distractor selection were significantly different from a random distribution based on a χ^2 test ($p < 0.001$).

	Option position				
	A	B	C	D	E
E is NOTA	2181 (18.9%)	2685 (23.3%)	3103 (26.9%)	2497 (21.6%)	1069 (9.3%)
E is Zero	1593 (15.6%)	2699 (26.4%)	2545 (24.9%)	1997 (19.5%)	1381 (13.5%)
Neither effect	9909 (20.5%)	11 504 (23.8%)	11 291 (23.4%)	8773 (18.2%)	6793 (14.1%)

selected preferentially over the other options for all three classes of questions. This pattern of distractor selection can be explained with a synthesis of the effects of primacy and middle bias. Middle bias accounts for a predisposition towards selecting options (b), (c), and (d) with an aversion to (a) and (e) while primacy would make the early distractors more likely to be selected. The results presented in Table V demonstrate a statistically significant position bias in students postinstruction as determined by a χ^2 test ($p < 0.001$ for each of the classes of questions). The failure to detect a zero bias in addition to the position bias may be a further indication that the instructor's efforts to confront zero bias were successful or may be a result of some of the zero answers forming common misconceptions in electricity and magnetism.

The distribution of correct answers was analyzed for the questions present in data set 2. The number of correct answers for each of the available options was tallied for the three classes of questions as shown in Table VI. The sum of all correct answers for each available option was examined to determine the overall trends of the professor's selection of correct answers. Table VI demonstrates a relatively uniform distribution of correct answers for problems in data set 2. As such, the position bias identified in Table V cannot be explained by students modifying their responses based on experience with the instructor.

The strength of these biases for position-based selection make the inclusion of well-vetted key balancing techniques valuable in the development of evaluation tools. A bias in the key either towards overly selected options or towards underselected options could result in either an increase or decrease in the average score.

E. FCI modified to introduce testwiseness effects

To examine the effect of parallel and opposite constructions and to further examine position bias, two modified versions of the FCI were created. These parallel and opposite constructions are options which are grammatically similar to preexisting options and have either similar or opposite meaning to the preexisting option. One modified version of the FCI (the FCI+) included 4 testwiseness treatments intended to increase student selection of the correct answers when utilizing testwiseness, while the second modified version (the FCI−) included 4 testwiseness treatments intended to decrease student selection of the correct answers when utilizing testwiseness. Each of these testwiseness treatments were used on 3 to 6 questions in their respective modified version of the FCI and each modified question was only affected by one of these treatments. These treatments are summarized in Table VII.

The results of applying these modified versions of the FCI were collected in data set 4. Students were divided into three approximately equal groups and given either the FCI+, FCI−, or an unmodified FCI as a post-test at the end of the semester. The average student score for each treatment was obtained by averaging the student score on all questions which had been modified by the treatment. To determine how each of these treatments affected student scores, the average student score was also obtained for each corresponding set of questions present in the unmodified FCI as a control. The difference between the modified average scores and unmodified average scores are presented in Table VIII. The difference in averages for each effect was small and for many treatments opposite to that which would have been expected from the

TABLE VI. Distribution of correct answers for 5 option multiple-choice questions from data set 2.

	Option position				
	A	B	C	D	E
E is NOTA	71 (29.7%)	39 (16.3%)	49 (20.5%)	39 (16.3%)	41 (17.2%)
E is Zero	88 (29.1%)	37 (12.3%)	75 (24.8%)	39 (12.9%)	63 (20.9%)
Neither effect	174 (12.8%)	276 (20.3%)	263 (19.4%)	367 (27.0%)	278 (20.5%)
Summed	333 (17.5%)	352 (18.5%)	387 (20.4%)	445 (23.4%)	382 (20.1%)

TABLE VII. Description of changes to FCI+ and FCI- instruments in data set 4.

Effect	Description
	FCI+
A+	Moved the correct answer from either A or E to C (4 questions).
B+	Moved the correct answer from either A or E to either B or D (4 questions).
C+	Moved paired distractors away from each other (made them non-consecutive) (3 questions).
D+	Added an additional distractor that is paired with the correct answer (6 questions).
	FCI−
A−	Moved the correct answer from either B, C, or D to A (4 questions).
B−	Moved the correct answer from either B, C, or D to E (4 questions).
C−	Moved a distractor paired with the correct answer away from the correct answer (made them non-consecutive) (3 questions).
D−	Added an additional distractor that is paired with a commonly selected distractor (5 questions).

testwiseness literature; many of the negative treatments produced positive increases in the average. This experiment supports the conclusion that the testwiseness effects explored, except for NOTA and zero bias, are weak effects postinstruction when the students have substantial content knowledge.

IV. DISCUSSION

This study investigated three research questions; these will be discussed in the order proposed.

To what extent do the instruments used in physics education research conform to well-established item writing rules? Many extensively researched instruments common to PER use distractors that may be preferentially avoided by students because of testwiseness, test taking strategies that do not require correct content knowledge. Table II shows some instruments with unbalanced keys, many instances of the NOTA option which has been identified in the literature as problematic [13,17,20], and still more instances of the zero option identified here as having potential testwiseness effects.

Are there testwiseness effects that are specific to scientific assessment instruments? The existence of NOTA bias identified in studies of nonscientific examinations [13,17,20] was confirmed as an effect in both quantitative and nonquantitative examinations of scientific understanding. Zero bias, while not as strong as NOTA bias in all cases, was identified as a testwiseness effect specific to fields requiring quantitative reasoning. While some part of zero and NOTA bias could be attributable to the application of misconceptions where the testwiseness-affected distractor does not represent the misconception, the effects were still substantial for students who do not report confidence in their answers and thus are not applying strongly held misconceptions.

Does student application of testwiseness affect the outcome of the PER instruments at the item or instrument level? An analysis of the overall effect of testwiseness effects from NOTA and zero options on the FCI and CSEM showed a small effect on final score which suggests that testwiseness is not a validity threat to the use of the overall instrument. Item-level testwiseness effects were more substantial and should be considered in any item-level analysis of problem difficulty. Overall, these results for NOTA options agree with the findings of Haladyna and Downing [17] that NOTA options should be avoided, and extends this assertion to a physics environment.

Both NOTA and zero options were weak distractors when compared to the other distractors in the studied instruments. This could result in increased difficulty on questions in which either of these options are the correct answer causing the misinterpretation of the scores on such problems. NOTA and zero aversion could also increase the likelihood of students randomly selecting the correct answer without use of the proper content knowledge when these options are used as distractors.

Analysis of data set 3 indicated that students appear to be affected by a combination of the effects of middle bias and primacy, with recency having little effect. This combination fully supports the work of Attali and Bar-Hillel [26] regarding middle bias. It is only partially supportive of the arguments of Blunch and Payne [27,28] regarding the importance of primacy and recency. Key balancing, as

TABLE VIII. Average scores are for a selection of 3 to 6 questions present on the FCI or one of the two modified versions of the FCI in data set 4. The modified versions of the FCI had the options changed for many of their questions to elicit testwiseness effects that were intended to either improve or diminish student performance.

	Effects							
	A+	B+	C+	D+	A−	B−	C−	D−
Modified FCI average score	73.2%	69.4%	58.9%	47.2%	57.8%	63.1%	52.5%	57.3%
Control FCI average score	71.3%	67.2%	57.5%	51.0%	52.0%	59.9%	50.0%	56.8%
Difference	1.9%	2.2%	1.3%	−3.7%	5.7%	3.1%	2.5%	0.4%

described by Bar-Hillel and Attali [14], can be used to reduce any unintended effects of position bias.

The identification of zero bias suggests that there may still be a number of science-specific elements of testwiseness that are yet to be explored. While it could not be examined in this study, it seems possible that students may also have an aversion to selecting other extreme values such as infinity or the limit does not exist.

V. IMPLICATION FOR INSTRUCTION

The analysis above suggests that the use of questions with NOTA or zero as one of the distractors produces an instrument with effectively fewer distractors which can change the possibility of the students selecting the correct answer by chance. More importantly, it is quite possible that the use of a zero or NOTA option as the correct option may increase the effective difficulty of the problem without changing the physical concept tested. With this analysis, NOTA options should be eliminated from multiple-choice instruments. While zero options cannot be eliminated, they should be used sparingly. Instructors may also consider explicitly confronting students' zero bias.

VI. IMPLICATION FOR RESEARCH

The above analysis also suggests several potential ways in which use of NOTA, zero, and potentially other testwiseness-affected options may impact research. While these analyses demonstrate that testwiseness effects are lessened postinstruction, the effect on the pretest could modify normalized gain results [49]. One should also consider the potential affect of the inclusion of NOTA or zero options on item level validity and reliability when developing an instrument for research purposes. The effect of the correct answer position should also be considered and key balancing should be used to mitigate the effects of position bias. Overall, these results suggest that NOTA and zero options should be avoided, when possible, in the development of new instruments and the evaluation of results obtained from existing instruments.

The effect of testwiseness on item response patterns could affect research methodologies that use item rather than test level data including factor analysis and item response theory. Further, the observation that students are using cognitive strategies unrelated to their physics knowledge to answer some conceptual questions makes the relation between assessment results and student knowledge more tenuous. Testwiseness, as explored in this research, may be only one of many testing or problem-solving strategies that affect the interpretation of the conceptual instruments used in PER.

VII. LIMITATIONS AND FUTURE WORK

The work presented examined only the selective aversion to certain incorrect answers in situations where the correct answer was unknown. This suggests there may also be an aversion to selecting certain correct answers even when the correct content knowledge is present. This would represent a substantially more serious threat to validity and would be clinically more important; this effect will be the focus of future research. The existence of other testwiseness effects beyond zero bias that are specific to quantitative disciplines should also be explored. These effects may be more important in mathematics than in physics because of the wider range of extreme values (∞, the limit does not exist) that are available.

This work presented one experimental study, Sec. III E, which showed that the more subtle testwiseness effects were not important postinstruction, but more experimental work is needed.

VIII. CONCLUSION

This paper supported the existence of NOTA bias, a student's preferential selection of distractors different than the none of the above distractor. Zero bias was identified as a weaker, but still substantial testwiseness effect. Students showed some position bias selecting the central items in the answer choice list preferentially and avoiding the last distractor. Many popular PER conceptual instruments contain questions with NOTA or zero distractors. Some instruments, notably the FCI, have substantially unbalanced answer keys where the distribution of correct answers is not uniform. The effect of options that include grammatically similar structure to other options were shown to be weak postinstruction, when students have substantial content knowledge. However, if a significant pretest effect exists that is not present in a post-test, this could modify the normalized gain. As such, testwiseness effects should be considered and minimized in evaluation construction. Testwiseness should also be considered in item-level analysis where items contain a testwiseness-affected distractor or when the correct answer may be influenced by testwiseness.

[1] S. M. Downing, Construct-irrelevant variance and flawed test questions: Do multiple-choice item-writing principles make any difference?, Acad. Med. **77**, S103 (2002).

[2] W. T. Rogers and D. Harley, An empirical comparison of three-and four-choice items and tests: Susceptibility to testwiseness and internal consistency reliability, Educ. Psychol. Meas. **59**, 234 (1999).

[3] J. K. Smith, Converging on correct answers: A peculiarity of multiple choice items, J. Educ. Measure. **19**, 211 (1982).

[4] K. Scouller, The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay, Higher Educ. **35**, 453 (1998).

[5] J. P. Dolly and K. S. Williams, Using test-taking strategies to maximize multiple-choice test scores, Educ. Psychol. Meas. **46**, 619 (1986).

[6] C. I. Chase, Relative length of option and response set in multiple choice items, Educ. Psychol. Meas. **24**, 861 (1964).

[7] W. Evans, Test wiseness: An examination of cue-using strategies, J. Exp. Educ. **52**, 141 (1984).

[8] J. Millman, C. H. Bishop, and R. Ebel, An analysis of test-wiseness, Educ. Psychol. Meas. **25**, 707 (1965).

[9] J. J. Diamond and W. J. Evans, An investigation of the cognitive correlates of test-wiseness, J. Educ. Measure. **9**, 145 (1972).

[10] R. E. Sarnacki, An examination of test-wiseness in the cognitive test domain, Rev. Educ. Res. **49**, 252 (1979).

[11] R. Thorndike, Testing the test: Reliability, J. Couns. Dev. **63**, 528 (1985).

[12] T. M. Haladyna, *Developing and Validating Multiple-Choice Test Items* (Routledge, New York, 2012).

[13] B. B. Frey, S. Petersen, L. M. Edwards, J. T. Pedrotti, and V. Peyton, Item-writing rules: Collective wisdom, Teach. Teach. Educ. **21**, 357 (2005).

[14] M. Bar-Hillel and Y. Attali, Seek whence: Answer sequences and their consequences in key-balanced multiple-choice tests, Am. Stat. **56**, 299 (2002).

[15] R. A. Ellsworth, P. Dunnell, and O. K. Duell, Multiple-choice test items: What are textbook authors telling teachers?, J. Educ. Res. **83**, 289 (1990).

[16] W. Weiten, Violation of selected item construction principles in educational measurement, J. Exp. Educ. **52**, 174 (1984).

[17] T. M. Haladyna and S. M. Downing, Validity of a taxonomy of multiple-choice item-writing rules, Appl. Meas. Educ. **2**, 51 (1989).

[18] D. E. Powers and D. A. Rock, Effects of coaching on SAT I: Reasoning test scores, J. Educ. Measure. **36**, 93 (1999).

[19] A. L. Dudycha and J. B. Carpenter, Effects of item format on item discrimination and difficulty, J. Appl. Psych. **58**, 116 (1973).

[20] S. L. Knowles and C. A. Welch, A meta-analytic review of item discrimination and difficulty in multiple-choice items using none-of-the-above, Educ. Psychol. Meas. **52**, 571 (1992).

[21] R. B. Frary, The none-of-the-above option: An empirical study, Appl. Meas. Educ. **4**, 115 (1991).

[22] A. T. Oluwafemiv and A. E. R. Ifedayo, A study of item response and anchor bias in economics objective tests among senior secondary school students in Osun state, Acad. J. Inter. Stud. **2**, 173 (2013).

[23] N. S. Fagley, Positional response bias in multiple-choice tests of learning: Its relation to testwiseness and guessing strategy, J. Educ. Psychol. **79**, 95 (1987).

[24] P. D. Jones and G. G. Kaufman, The differential formation of response sets by specific determiners, Educ. Psychol. Meas. **35**, 821 (1975).

[25] L. M. Stough, Research on multiple-choice questions: Implications for strategy instruction, in *71st Annual Convention of the Council for Exceptional Children*, San Antonio, TX, April 1993 (Council for Exceptional Children, Arlington, VA, 1993), pp. 1–11.

[26] Y. Attali and M. Bar-Hillel, Guess where: The position of correct answers in multiple-choice test items as a psychometric variable, J. Educ. Measure. **40**, 109 (2003).

[27] N. J. Blunch, Position bias in multiple-choice questions, J. Market. Res. **21**, 216 (1984).

[28] S. Payne, *The Art of Asking Questions: Studies in Public Opinion* (Princeton University Press, Princeton, NJ, 2014), Vol. 3.

[29] F. M. Carp, Position effects on interview responses, J. Gerontol. **29**, 581 (1974).

[30] Jason Millman, *How to Take Tests* (McGraw-Hill, New York, 1969).

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

[32] R. K. Thornton and D. R. Sokoloff, Assessing student learning of Newton's laws: The force and motion conceptual evaluation and the evaluation of active learning laboratory and lecture curricula, Am. J. Phys. **66**, 338 (1998).

[33] C. Singh and D. Rosengrant, Multiple-choice test of energy and momentum concepts, Am. J. Phys. **71**, 607 (2003).

[34] L. G. Rimoldini and C. Singh, Student understanding of rotational and rolling motion concepts, Phys. Rev. ST Phys. Educ. Res. **1**, 010102 (2005).

[35] L. Ding, Ph.D. thesis, North Carolina State University, 2007.

[36] D. Hestenes and M. Wells, A mechanics baseline test, Phys. Teach. **30**, 159 (1992).

[37] D. P. Maloney, T. L. O'Kuma, C. J. Hieggelke, and A. Van Heuvelen, Surveying students' conceptual knowledge of electricity and magnetism, Am. J. Phys. **69**, S12 (2001).

[38] P. V. Engelhardt and R. J. Beichner, Students' understanding of direct current resistive electrical circuits, Am. J. Phys. **72**, 98 (2004).

[39] L. Ding, R. Chabay, B. Sherwood, and R. Beichner, Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment, Phys. Rev. ST Phys. Educ. Res. **2**, 010105 (2006).

[40] J. Li and C. Singh, Developing a magnetism conceptual survey and assessing gender differences in student understanding of magnetism, AIP Conf. Proc. **1413**, 43 (2012).

[41] B. R. Wilcox and S. J. Pollock, Coupled multiple-response versus free-response conceptual assessment: An example from upper-division physics, Phys. Rev. ST Phys. Educ. Res. **10**, 020124 (2014).

[42] A. Tongchai, M. D. Sharma, I. D. Johnston, K. Arayathanitkul, and C. Soankwan, Developing, evaluating and demonstrating the use of a conceptual survey in mechanical waves, Int. J. Sci. Educ. **31**, 2437 (2009).

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

[44] J. Stewart and G. Stewart, Correcting the normalized gain for guessing, Phys. Teach. **48**, 194 (2010).

[45] R. Leinonen, M. A. Asikainen, and P. E. Hirvonen, Overcoming students' misconceptions concerning thermal physics with the aid of hints and peer interaction during a lecture course, Phys. Rev. ST Phys. Educ. Res. **9**, 020112 (2013).

[46] L. Bao and E. F. Redish, Model analysis: Representing and assessing the dynamics of student learning, Phys. Rev. ST Phys. Educ. Res. **2**, 010103 (2006).

[47] J. Stewart, M. Miller, C. Audo, and G. Stewart, Using cluster analysis to identify patterns in students' responses to contextually different conceptual problems, Phys. Rev. ST Phys. Educ. Res. **8**, 020112 (2012).

[48] J. Stewart, H. Griffin, and G. Stewart, Context sensitivity in the Force Concept Inventory, Phys. Rev. ST Phys. Educ. Res. **3**, 010102 (2007).

[49] R. R. Hake, Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses, Am. J. Phys. **66**, 64 (1998).