



Examining Preservice Elementary Teachers' Answer Changing Behavior on a Content Knowledge for Teaching Science Assessment

Katherine E. Castellano¹ · Jamie N. Mikeska¹ · Jung Aa Moon² · Steven Holtzman¹ · Jie Gao¹ · Yang Jiang¹

Accepted: 11 May 2022 / Published online: 23 June 2022
© The Author(s), under exclusive licence to Springer Nature B.V. 2022

Abstract

Preservice elementary teachers (PSTs) prepare for various standardized assessments, such as the *Praxis*® licensure assessment. However, there is little research on test-taking behavior and test-taking strategies for this examinee population. A common belief and instruction given in some test preparation materials is that examinees should stick to their initial answer choice. Decades of research has debunked this belief, finding that generally examinees benefit from answer changing behavior. However, there is minimal research on answer changing behavior among PSTs. Moreover, there is little research examining answer changing behavior for tests assessing constructs that integrate content and practice, or across different technology-enhanced item types. We use an online Content Knowledge for Teaching (CKT) assessment that measures PSTs' CKT in one science area: matter and its interactions. In this study, we analyzed process data from administering the online CKT matter assessment to 822 PSTs from across the US to better understand PSTs' behaviors and interactions on this computer-based science assessment. Consistent with prior research findings, this study showed that examinees who changed their responses benefited more often than were harmed by doing so with higher-performing examinees benefiting more than lower-performing examinees, on average. These findings also were consistent across item types. Implications for computer-based CKT science assessment design and delivery are discussed.

Keywords Answer changing · Process data · Technology-enhanced items · Content knowledge for teaching science · Matter and its interactions

Research on answer changing behavior during testing dates back to as early as 1929 (Mathews, 1929), where more than 53% of the answers changed were from wrong to right in an educational psychology course test, and more points were gained than lost as a result. Since then, many more studies have showed similar results (e.g., Al-Hamly & Coombe, 2005; Waddell & Blankenship, 1994), refuting the common belief and instruction given in some test preparation materials that examinees should stick to their initial answer choice. Research studies have also investigated the relationship between different variables and answer changing behaviors, such as gender (e.g., Bath, 1967; Geiger, 1991a), ethnicity (Payne, 1984), item difficulty and type of content being assessed in the item (Geiger, 1991b; Ramsey et al.,

1987; Vidler & Hansen, 1980), students' metacognition (McConnell et al., 2012), and test anxiety (Green, 1981). Despite the depth of research on answer changing behavior, there is no or little research on answer changing behavior for the preservice teacher (PST) test-taker population, for tests assessing constructs that integrate both content and practice, or across different technology-enhanced item types.

This descriptive study addresses these three gaps to begin to build evidence about the extent to which historical findings on answer changing apply to a different test-taker population, a more complex construct integrating science content and teaching practice, and increasingly used item types. Integrated constructs of science content knowledge and specific practices (e.g., teaching or science practices) are on the rise and include, for instance, content knowledge for teaching (CKT) and performance expectations from the Next Generation Science Standards (NGSS). A better understanding of answer changing behavior for this population of test-takers, constructs, and item types can help develop a more nuanced understanding of the ways in which the PST population

✉ Jamie N. Mikeska
jmikeska@ets.org

¹ ETS: Educational Testing Service, 660 Rosedale Road, Princeton, NJ 08541, USA

² Honeywell International Inc, 2555 Smallman St, Suite 200, Pittsburgh, PA 15222, USA

interacts with increasingly complex items and provide guidance on test-taking strategies.

In this study, we used a research instrument, which was developed through a grant funded by the National Science Foundation (NSF), designed to assess elementary PSTs' CKT in one high-leverage science content area: matter and its interactions. The CKT matter assessment includes various technology-enhanced items (TEIs) as well as traditional single select multiple choice items. We administered the test to 822 PSTs across the United States and used the answer response process and selection data to address the following research questions (RQs):

1. How does PSTs' answer changing behavior relate to item-level and total test scores on the CKT matter test?
2. To what extent does answer changing behavior differ across PSTs' performance level?

Before addressing these research questions, we provide more background on PST assessments and CKT and answer changing behavior research.

Background

Preservice Teacher Assessments and CKT

Historically, assessments in teacher education have focused on measuring teachers' understanding of the subject matter that they will be responsible for helping their own K-12 students master (Wilson, 2016). For example, many states have adopted and currently use subject matter knowledge assessments, such as ETS's *Praxis*® series, to determine if teacher candidates in their states are ready to enter the teaching profession. While an understanding of the student-level content domain has been shown as critical for both novice and experienced teachers, research has suggested that teachers need to know how to use their subject matter knowledge to engage in specific teaching practices (Ball et al., 2008; National Academies, 2015; National Research Council, 2013; Shulman, 1986).

Ball et al. (2008) coined the term “content knowledge for teaching” (CKT) to refer to the content knowledge that teachers need to use as they engage in the work of teaching in specific disciplines. In science teacher education, there has been a concerted effort to develop assessments of the more practice-based aspects of science teachers' CKT—their pedagogical content knowledge—and the ways that they use this knowledge in the work of teaching science. These CKT assessments have included analysis of video clips, interviews, observational protocols, performance assessments, and in-take surveys to assess this aspect of science teachers'

CKT (Bertram & Loughran, 2012; Henze & van Driel, 2015; Park & Oliver, 2008; Park & Suh, 2015; Roth et al., 2011).

However, most of these assessments require an extensive amount of time to both administer and score. The field is in dire need of CKT science assessments that can be administered efficiently and effectively on a large scale. Such assessments would support the field in providing ongoing feedback to teacher educators and professional development facilitators, fostering a science teacher workforce equipped to support student learning and providing a way for researchers to investigate variability within and across programs and sites.

To that end, a NSF-funded research team developed a new CKT assessment to measure PSTs' CKT in one high-leverage science content area: matter and its interactions (see, for instance, cktscience.org and Mikeska & Castellano, 2021). This assessment is one of the first of its kind, as it uses automatically scored assessment items targeting the teachers' CKT in this area. But to fully understand what these novel CKT science assessments measure and how examinees interact with them, the field needs to examine both examinee response and process data and their interactions. Process data refers to information about respondents' interactions with computer-based tasks that are captured via log files to show the specific actions taken when engaging with an item. In this study, we captured PSTs' time-stamped clicks as they responded to each CKT matter assessment item online and examined this process data to better understand the PSTs' answer changing behaviors and how they relate to their performance levels. We focus initially on their answer changing behavior, as that has been a major area of research on test-takers' behaviors dating back for almost 100 years. In particular, this study focuses on determining if the same patterns in answer changing behavior that prior research has noted for other assessments and other test-taking populations are similar for this new construct (CKT), a different test-taker population (elementary PSTs), and innovative item types.

Answer Changing Behavior

In a review of 56 studies since 1929, Al-Hamly and Coombe (2005) reported empirical results consistently supported four key findings on answer changing behavior: (1) a minority of item responses are changed, (2) a majority of examinees make at least one change, (3) answer changers tend to make more wrong-to-right (WR) than right-to-wrong (RW) changes, and (4) more answer changers reap score gains than score losses. We briefly review these findings.

Studies on answer changing behaviors have found that the percentage of item by examinee responses being changed ranges from about 2 to 10%, with 60 to 96% of examinees making at least one change. That is, most examinees make at least one change, but, on average, they do not change answers to many items. For example, Al-Hamly and Coombe

(2005) administered the Michigan English Language Institute College English Test (MELICET) to 286 college students and found that 62% of students changed their answers, while only 2.65% of test answers over all item by examinee responses were changed. Similar findings have been shown across various examinee populations and using different test content, including eleventh grade mathematics test-takers (Jeon et al., 2017), medical students taking a medical assessment (Bauer et al., 2007; Ouyang et al., 2019), college students in the field of education completing an educational psychology final course test (Stylianou-Georgiou & Papanastasiou, 2017), community college course exams (Freidman-Erickson, 1994), and examinees taking the GRE General Test (Liu et al., 2015). While the prevalence of RW and WR changes varies across studies, generally studies find at least double the amount of WR to RW changes, indicating that answer changes to items are twice as likely to be beneficial rather than detrimental changes (Al-Hamly & Coombe, 2005; Bauer et al., 2007; Freidman-Erickson, 1994; Liu et al., 2015; Stylianou-Georgiou & Papanastasiou, 2017; Waddell & Blankenship, 1994).

Given the generally higher rates of WR than RW item score changes, more examinees have a net gain in their total test score than a net loss from answer changing. For example, of all the students who changed answers in the Al-Hamly and Coombe (2005) study, 19% had a net loss and 57%, or three times more, had a net gain. Similar results have been found in other studies using results from the GRE and an elementary mathematics assessment (Bauer et al., 2007; Bridgeman, 2012; Liu et al., 2015; Van der Linden et al., 2012), suggesting that it can be far more likely for an examinee to benefit from changing answers than to be disadvantaged by it.

Given the depth of literature supporting these findings, we predicted (RQ1) that similar patterns would also hold for our assessment of an integrated construct (CKT of matter and its interactions), for a specialized population (PSTs), and using varied item types (single select multiple choice and two TEI types). We predicted there would be some variation by our item types of interest, but that these findings would generally hold for each item type as well. Our empirical study set out to investigate these predictions.

Finally, answer changing behavior has been hypothesized to be related to test-takers' performance level. For example, research studies have shown that higher scoring students tend to make fewer answer changes—both overall (Liu et al., 2015) and for WW answer changes (Al-Hamly & Coombe, 2005). In addition, research has shown that higher performing examinees also tend to experience the greatest benefit (i.e., more WR than RW changes on items and more score gains than score losses) when changing their answer responses compared to lower ability examinees (Jeon et al., 2017; Liu et al., 2015). Based on these previous findings, in our study, we predicted that higher performing examinees would benefit more from

answer changing than lower performing examinees regardless of item type (RQ2).

Data Description

Sample

We administered the field test of the CKT assessment on matter and its interactions to PSTs currently enrolled in a teacher preparation program for elementary education in the US. We recruited the PSTs from the pool of *Praxis*® Elementary Science (Test 5005) test-takers who completed the test from January 2018 to June 2019.¹ This *Praxis*® Science test is part of the Elementary Education: Multiple Subjects assessment and is used to ensure that PSTs possess the science content knowledge necessary to enter the teaching profession at the elementary level. Many states require passing scores on this test for PSTs to receive a generalist elementary teaching license.

We used a stratified random sample with strata defined by gender (female/male), ethnicity (white/not white), geographical region (Northeast/Midwest/South/West), and *Praxis*® Science score quartile (Q1, Q2, Q3, Q4). We fully crossed these four strata to define $2 \times 2 \times 4 \times 4 = 64$ unique cells from which we randomly sampled PSTs. Of the 960 selected PSTs, 822 completed our assessment. These 822 were largely representative of the full *Praxis*® Science test-taking population with all sample percentages within 4% of the target population.²

CKT Matter Instrument

An extensive item development process was used to assemble the CKT matter assessment field test form. Each CKT matter item was aligned to one of five sub-content areas (e.g., changes in matter) and one of seven Work of Teaching Science (WOTS) instructional tools (e.g., scientific explanations). Items were developed in three batches over an 18-month period with each batch being reviewed and revised through external expert reviews and cognitive interviews with PSTs and current elementary school teachers (see Figs. 1, 2 and 3 for sample items). Approved items were piloted with a sample of about 200 PSTs from *Praxis*®

¹ We excluded any *Praxis*® Elementary Science test-takers who participated in our earlier CKT matter item pilots.

² The sample was 92% female (pop=93%), 80% White (pop=80%), 4% Midwest (pop=4%), 22% Northeast (pop=22%), 46% South (pop=49%), 28% West (pop=25%), 22% in the first quartile (Q1) of the *Praxis*® Science score distribution (pop=26%), 27% in Q2 (pop=27%), 24% in Q3 (pop=23%), and 27% in Q4 (pop=24%).

Fig. 1 Example of single-selection multiple-choice (SSMC) CKT matter item

<p>Ms. Jones had her fifth-grade students sort materials by their properties. One tool given to her students to help in this process was a magnet. The materials tested included the following.</p> <ul style="list-style-type: none"> ▪ Gray iron bar ▪ Block of wood ▪ Glass marble ▪ Button magnet ▪ Candle ▪ Salt ▪ Water ▪ Bar magnet ▪ Paper towel ▪ Plastic fork ▪ Steel nail ▪ Silver crayon
<p>Given the materials supplied by Ms. Jones, which of the following misconceptions could be reinforced through this activity?</p> <p>A) All solid objects are magnetic. B) All metal objects are magnetic. C) All silver-colored objects are magnetic. D) Only objects with “magnet” in their name are magnetic.</p>

Elementary Science (5005) test-takers to obtain preliminary item statistics.³

We reviewed pilot data to identify items that should be dropped for further consideration or revised. We used the final selected and revised items to assemble our 60-item field test form in accordance with our test blueprint—proportions of items by five sub-content areas and seven WOTS instructional tool categories. (See supplemental Online Appendix A for details).

Item Types

The 60-item field test also contained a variety of item types: single-select multiple choice (SSMC) ($n = 24$), multiple-select multiple choice (MSMC) ($n = 17$), grid multiple selection (one-per-row) ($n = 7$), match multiple selection ($n = 5$ items), inline choice multiple selection ($n = 5$), and grid multiple selection (varying-per-row) ($n = 2$). For this study, we focused on the three item types with more than 5 items, resulting in a total of 48 items. Figures 1, 2 and 3 provide examples of the three item types of interest using items from our pool that were not on the field test (to preserve the security of those items).

As shown in Fig. 1, SSMC items involve selecting a single choice among a set of options and a single point is earned if this selection is correct. In contrast, MSMC items involve selecting more than one option among a set of options. As shown in Fig. 2, for each MSMC item that we consider in this study, the number of required selections for a correct response

is specified in the item stem. Given that examinees know how many selections they are required to make for a correct response, a single point is earned if all those selections are correct. There is no partial credit for having a subset of the selections correct. Finally, grid multiple selection (one-per-row), which we refer to simply as “grid items,” involve making one selection per row of a grid, as shown in Fig. 3. Similar to MSMC items, a single point is awarded if all rows have correct selections. No partial credit is awarded if a subset of the selections is correct. There are various methods for writing and scoring MSMC and grid items, but we followed the practices used for the *Praxis*® Elementary Science assessment which specifies the number of required selections for MSMC items and uses dichotomous scoring rules for MSMC and grid items.

The items for each of the three item types covered a range of material (by content sub-topic and WOTS categories) and ranged in difficulty and discrimination (item-total score correlation), but these content and statistical specifications of the items were not controlled to match exactly. See Supplemental Online Appendix A for more details on content coverage by item type and the extent that distributions of item statistics (item difficulty, item-total score correlation, and item time) varied by item type (Figure A1). Although items of each type differ on more than format, some of the differences in item characteristics may be partly due to the format. For instance, certain content may be better assessed with some item formats than others. Similarly, the multi-part response structure for grid items may lead to higher difficulty and require more time than SSMC items.

Instrument Delivery and Process Data Capture

The CKT matter assessment was delivered online with every selection automatically recorded. That is, any change to a response selection was recorded in the log data for each PST’s

³ Pilot recruitment occurred at the beginning of 2018 so only included *Praxis*® Elementary Science test-takers in the 2018 calendar year, whereas Field Test recruitment occurred in July 2019 so it included *Praxis*® Elementary Science examinees from January 2018 to June 2019. PSTs in the pilot were ineligible to participate in the field test.

Fig. 2 Example of multiple-selection multiple-choice (MSMC) CKT matter item

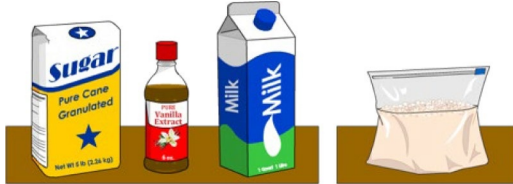
<p>In Ms. Quintana’s second-grade class, students explore the properties of different solids and liquids. Based on the exploration findings, students create definitions for solids and liquids.</p> <p>While completing the definition for liquids, one student makes the claim that “all substances that look like they take the shape of their containers are liquids.” Ms. Quintana is planning to include a follow-up activity for students to collect more data and refine their ideas.</p>
<p>Which TWO of the following materials will best challenge the claim and help the student improve his or her definition?</p>
<p>A) Maple syrup B) Ice block C) Salt D) Milk E) Rice</p>

test session. Examinees were not allowed to return to previously answered items so all answer changes were for the first and only encounter with the item. In addition, examinees could not advance on the test until a complete response was provided for each item. For instance, for MSMC items, if the item specified that three options should be selected, the examinee could

not move forward until they had made three selections. Similarly, for grid items, examinees could not move forward until a single selection had been made for each row.

Examinees could change their response as many times as they would like when interacting with an item. Examinees changed their responses for a particular item from 0 to 37 times, but the

Fig. 3 Example of a grid CKT matter item

<p>Ms. Henderson’s fifth-grade class is making ice cream using milk, sugar, and vanilla flavoring. After mixing the ingredients in a baggie, the students observe the properties of the mixture, and are then asked to describe what happens to the sugar and vanilla in the milk.</p>		
		
<p>Nico says that there is no sugar or vanilla in the mixture because he cannot see them in the milk. Ms. Henderson is considering several instructional moves to help develop Nico’s conceptual understanding of what happens to the sugar and vanilla.</p>		
<p>For each instructional move, select the appropriate cell in the table to indicate whether or not the move will help develop Nico’s conceptual understanding.</p>		
Instructional Move	Will Help Develop Nico’s Conceptual Understanding	Will Not Help Develop Nico’s Conceptual Understanding
<p>Mix one teaspoon of vanilla with one tablespoon of water and one teaspoon of vanilla with one tablespoon of milk. Have Nico observe both mixtures and describe any changes to the color, volume, and consistency of the water and milk.</p>	<input type="checkbox"/>	<input type="checkbox"/>
<p>Weigh the recipe amounts of milk, sugar, and vanilla before mixing, and then weigh the mixture of the ingredients. Have Nico compare the sum of weights of each ingredient with the weight of the mixture to show that the sugar and vanilla are present in the milk mixture.</p>	<input type="checkbox"/>	<input type="checkbox"/>
<p>Dissolve one tablespoon of sugar in one cup of water and then evaporate the water so the sugar recrystallizes. Have Nico explain what happens to the sugar when it is mixed with water based on his observations.</p>	<input type="checkbox"/>	<input type="checkbox"/>
<p>Cool one cup of milk and heat another cup of milk before adding sugar and vanilla. Have Nico make observations of the mixtures of the sugar and vanilla in the cold and warm milk, and then have Nico record his observations in a chart for comparison.</p>	<input type="checkbox"/>	<input type="checkbox"/>

mean and median number of changes when a change was made were only 1.5 and 1.0, respectively. For this study, we focus on the first and last responses. The first response was defined as the first set of selections that satisfied the required selections for the item. For instance, if a MSMC item stated that two responses should be selected, and an examinee selected option A, unselected option A, selected option B, and then selected option C, the examinee's first response would be options A and B as they were the first two options selected (even though option A was unselected before selecting option B). The last or final response for an examinee was the examinee's submitted response. In this example, the examinee's final response would be options B and C. In some cases, the first and last response may be the same even if the examinee made changes in between these selections. We still consider such cases as involving an answer change, but there would be no score change. Accordingly, for each item, we have six different types of possible response patterns denoted by "R" for "right" (correct) and "W" for "wrong" (incorrect) as is the tradition in answer change literature: R—no answer changes and item response is right, W—no answer changes and item response is wrong, RR—examinees make answer changes but the first and final responses are the same right selection, WW—examinees make answer changes but the first and final responses are either the same wrong selection or two different wrong selections, RW—examinees change their initial right selection to a final wrong selection, WR—examinees change their initial wrong selection to a final right selection. For this study, we focus on the last four patterns that involve answer changing.

Findings

Research Question 1: How Does PSTs' Answer Changing Behavior Relate to Item-Level and Total Test Scores on the CKT Matter Test?

To address the first research question, we first focus on the item level answer change results and then the impact on examinee scores.

Item Level

Every item had some examinees making answer changes, but across the items, the percentage of examinees making response changes overall and by each of the four answer change response patterns varied. For almost all items, regardless of item type and consistent with the literature, the percentage of examinees who made WR changes exceeded that of those who made RW changes. But the extent that they differed varied by item with differences (WR – RW) ranging from –3 to +21 percentage points and a mean of 5.7 percentage points.

The first row in Table 1 shows that for the full set of 48 items, on average, 19% of examinees made changes to an item, or, equivalently since all examinees completed the same number of items, 19% of all ($48 \times 822 = 39,456$) item-by-examinee responses were changed. Across the item types, this percentage ranged from 15.9% for SSMC items to 25.5% for grid items with more answer changes, on average, for the two TEIs. Across all item types, the overall average percentage making RW changes per item, or detrimental changes, was 2.7% with little variation across item types (2.5 to 2.8%). The overall average percentage making WR changes per item, or beneficial changes, was higher at 8.4%. The average WR rates across item types was also similar, ranging from 7.7 to 9.1% with SSMC items having the highest average percentage and MSMC the lowest. Accordingly, the "item gain-to-loss ratio" as defined in Liu et al. (2015) as the percentage of beneficial changes (WR) to detrimental (RW) changes is also similar (2.9 to 3.3) and close to the overall average of 3.1. That is for each item type (and over all item types), on average, there are about 3 times as many examinees making beneficial changes as those making detrimental changes on each item.

Table 1 also reveals that for MSMC and grid items, on average, a noticeable proportion of PSTs made WW changes (8% and 13%, respectively), whereas only about 3% of PSTs made such changes for SSMC items, on average. The somewhat high rates of examinees making WW changes for

Table 1 Percentage of examinees making each type of response pattern for an item averaged over all items of each item type with standard deviations in parentheses

	All Items (48 items)	SSMC (24 items)	MSMC (17 items)	Grid (7 items)
Mean percent (SD) of examinees changing a response on an item	19.0 (6.5)	15.9 (5.9)	20.8 (5.5)	25.5 (4.5)
RW	2.7 (1.7)	2.8 (1.9)	2.7 (1.6)	2.5 (1.1)
WR	8.4 (4.1)	9.1 (4.4)	7.7 (4.1)	8.0 (2.7)
WW	5.9 (4.5)	2.6 (1.5)	7.6 (2.8)	13.1 (3.9)
RR	2.0 (1.4)	1.5 (1.1)	2.8 (1.5)	1.9 (1.2)
Item gain-to-loss ratio	3.1	3.3	2.9	3.2

Note. RW right-to-wrong, WR wrong-to-right, WW wrong-to-wrong, RR right-to-right, SSMC single-selection multiple-choice items, MSMC multiple-selection multiple-choice items

MSMC and grid items may be attributed to the fact that these item types require multiple selections. It is thus easier to make changes and still result in an incorrect response. See the “Discussion” section for further consideration of this result.

Examinee Level

At the examinee level, we reviewed the impact of answer changes on examinees’ overall test scores. As shown in the first row of Table 2, across all 48 items, an overwhelming majority of examinees (821 out of 822) made at least one change. Similarly, very high proportions of examinees made at least one change to the 24 SSMC or 17 MSMC items (94 and 95%, respectively). For grid items, “only” 81% of all examinees made at least one change over the 7 grid items. However, this lower proportion may be more a consequence of the fewer opportunities to make a change (with only 7

items) than indicating a lower likelihood that examinees will change grid items; if we randomly select only seven SSMC or MSMC items, we also find lower proportions of examinees making at least one change.

Examinees who made at least one answer change (or “answer changers”), tended to change more than one item. Over all 48 items, answer changers changed 9 items or 19% of the items on average. Similarly, answer changers for SSMC or MSMC items, on average, changed about 4 items, or 17% and 22% of the items, respectively. Answer changers for grid items changed about 2 of the 7 items on average, or 31% of the items.

Answer changers, on average, made score gains regardless of item type. Overall, examinees with answer changes over the 48 items ($N=821$) gained an average of 2.73 points, or 6% of the total possible 48 points. For each item type, answer changers, on average, made slight gains in their total test score after changing: for SSMC items, answer

Table 2 Impact of answer changing on examinee test scores

	All Items (48 items)	SSMC (24 items)	MSMC (17 items)	Grid (7 items)
Examinees who made at least one answer change (answer changers)				
Total number of examinees	822	822	822	822
Number of answer changers	821	770	784	665
Percentage examinees	100%	94%	95%	81%
Mean number of items changed (SD)	9.13 (4.58)	4.07 (2.49)	3.70 (2.04)	2.20 (1.13)
Mean score change (SD)	2.74 (2.86)	1.62 (1.99)	0.88 (1.38)	0.47 (0.96)
Examinees who made at least one answer change with score gains				
Number of answer changers	648	562	455	310
Percentage of answer changers	79%	73%	58%	47%
Mean number of items changed (SD)	9.68 (4.60)	4.32 (2.60)	4.09 (2.00)	2.37 (1.15)
Mean score change (SD)	3.62 (2.53)	2.44 (1.63)	1.78 (1.02)	1.32 (0.58)
Examinees who made at least one answer change with score losses				
Number of answer changers	67	95	102	90
Percentage of answer changers	8%	12%	13%	14%
Mean number of items changed (SD)	7.97 (3.58)	3.62 (2.30)	3.47 (1.89)	2.32 (1.09)
Mean score change (SD)	-1.52 (0.77)	-1.37 (0.68)	-1.21 (0.43)	-1.07 (0.25)
Examinees who made at least one answer change with scores unchanged				
Number of answer changers	106	113	227	265
Percentage of answer changers	13%	15%	29%	40%
Mean number of items changed (SD)	6.51 (3.97)	3.21 (1.79)	3.00 (1.98)	1.96 (1.06)
Mean score change (SD)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
Examinee gain-to-loss-ratio	9.7	5.9	4.5	3.4

Note that for the “All Item Types” column, the score changes are across all 48 items, but for each of the item type columns, score changes are just for items of that type. For instance, a test-taker could increase their score overall and thus be classified as a score gainer for all items but could lose points for a particular type of item, say, grid items, and thus for grid items, would be considered a test-taker with a score loss. Note caution should be used when comparing the mean numbers of items changed or the mean score change across the three item types given each item type has a different number of items overall. Mean score change is the average difference in total score points from initial to final response. For instance, score gainers, gained an average of 4.32 points from their initial to final responses on SSMC items

Note. SSMC single-selection multiple-choice items, MSMC multiple-selection multiple-choice items

changers ($N=770$) gained an average of 1.62 points, or 7% of the possible 24 points for SSMC items; for MSMC items, answer changers ($N=784$) gained 0.88 points (5% of 17 points); and for grid items, answer changers ($N=665$) gained 0.47 points (7% of 7 points).

Table 2 also breaks down answer changers into three groups: those with score gains, score losses, and no score change. If the impact of answer changing was random, then examinees would be equally likely to lose points, gain points, or have an unchanged total score after changing. Separate chi-squared tests by item type comparing the observed distribution to this expected uniform distribution reveal significant results indicating that the distributions are significantly different than uniform (SSMC: $P(\chi^2(2) > 545.47) \approx 0$; MSMC: $P(\chi^2(2) > 245.18) \approx 0$; Grid: $P(\chi^2(2) > 121.88) \approx 0$). This result is not surprising given most answer changers gain points for SSMC (73% of answer changers) and MSMC (58% of answer changers) items and about half (47% of answer changers) are gainers for grid items compared to only 12 to 14% of answer changers losing points after answer changing. Post hoc pairwise comparisons within item type (with Bonferroni corrections for multiple comparisons) reveal that for each item type, there are significantly more total score gainers than losers ($p \approx 0$ in each case). For SSMC and MSMC items, there are also significantly more answer changers who gain points than have the same score ($p \approx 0$ in each case), and for MSMC and grid items, there are significantly more answer changers with scores unchanged than with score losses ($p \approx 0$ in each case). Accordingly, PSTs tended to either come out ahead or no worse off if they changed some of their responses throughout the test.

Comparing the percentage of answer changers with gains to those with losses, we obtain the examinee gain-to-loss ratio in the bottom row of Table 2 (as defined in Liu et al. (2015) as the percentage of examinees with score gains to those with losses); over all items, almost 10 times as many of the answer changers had total test score gains than score losses. For SSMC items, about 6 times as many answer changers had test score gains than score losses compared to about 4.5 as many having score gains for MSMC items and 3.4 times as many for grid items. In all cases, there were substantially more answer changers whose scores increased than decreased, indicating that answer changing regardless of item type tends to be more beneficial than detrimental to an examinee's test score.

Moreover, the magnitude of the score change is, on average, higher for those with score gains than losses. T -tests comparing the score gain to (absolute value) score loss are significant for each item type (SSMC: $P(t(311) > 10.96) = 0$; MSMC: $P(t(382) > 9.00) = 0$; Grid: $P(t(342) > 6.06) = 0$). For instance, for SSMC items, answer changers with total test score gains, on average, gained 2.44 score points, whereas those with score losses, on average, lost only 1.37 score points. Thus, examinees disadvantaged by their answer

changing behavior tended to be disadvantaged less than examinees advantaged by their answer changing behavior were advantaged.

Research Question 2: To What Extent Does Answer Changing Behavior Differ Across Preservice Teachers' Performance Level?

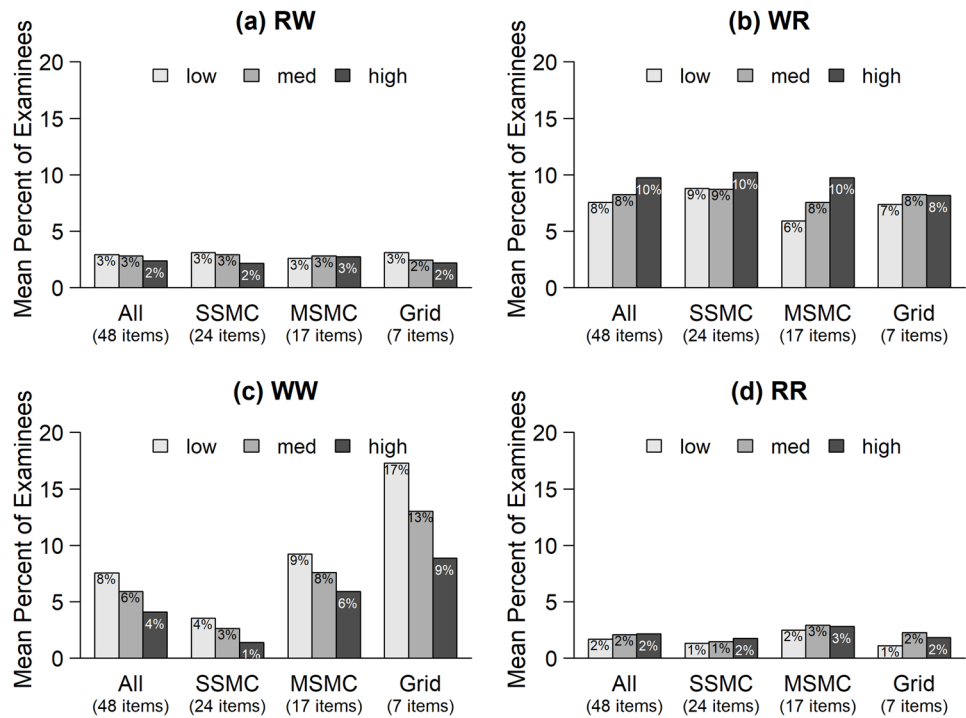
We further explore answer change behavior by item type as a function of examinee performance level to examine the extent that the positive relationship between examinee performance and benefit of answer changing holds for our population, construct, and item types. We classified examinees into performance categories by their performance on an external measure of science teaching ability—their *Praxis*® Science scores. The *Praxis*® Science test assesses a broader content domain (elementary science) than our instrument (matter and its interactions) and focuses more on pure content knowledge than CKT, but PSTs were highly motivated to perform well on the *Praxis*® test as it assists in obtaining their teaching license and it is moderately correlated with our instrument ($r=0.53$; correlation disattenuated for measurement error on both instruments = 0.64). Unlike the total score for the CKT assessment, which reflects examinees' gains or losses from answer changing observed in this study, the *Praxis*® score is an independent, external measure. We categorized PSTs into three ability groups: low (bottom *Praxis*® Science quartile), medium (middle 50%), and high (top quartile). As with research question 1, we first review results at the item level and then turn to the impact on examinee scores.

Item Level

For all 48 items, the mean percentages of examinees making a change to an item were all within about 1.5 percentage points across ability groups with 19.7% of low-performing PSTs to 18.3% of high-performing PSTs making changes to an item, on average (see Supplemental Online Appendix Table B1 for the full results). Similarly, on average, 16.8% of low-performing PSTs made a change to a given SSMC item compared to 15.5% of high-performing PSTs. For MSMC items, the percentage of examinees making answer changes, on average, per item were about 21% (20.2%, 20.9%, and 21.3% for low, medium, and high performing, respectively). For grid items, there was more variation across performance levels with 28.8% of low-performing examinees making changes on average to a given grid item compared to 25.9% and 21.0% for medium and high ability levels, respectively. Accordingly, for a given grid item, the bottom 25% of examinees were 1.4 times more likely to change a response on average than the top 25% of examinees.

The four panels of Fig. 4 illustrate the average percentage of examinees per item making each of the four answer

Fig. 4 The percentage of examinees making each response pattern for an item averaged over items for each ability level



Note. SSMC single-selection multiple-choice items; MSMC multiple-selection multiple-choice items. RW right-to-wrong; WR wrong-to-right; WW wrong-to-wrong; RR right-to-right.

change response patterns averaged over all items of each item type. Visual inspection of these panels reveals trends by ability level for each answer change response type and item type. However, to understand better the relationships between the tendency of making make each type of answer change by ability level within each item type, we fit multinomial logistic regression models—one for each item type. We used the person by item data for all responses in which examinees made an answer change. For instance, if an examinee changed responses to five items, their data will appear in five rows—one for each item with a response change. The type of answer change response type (RW, WR, RR, WW) was the dependent variable with RW as the reference category, and the PST ability groups were independent variables (with low-performing as the reference category). We also included indicators for each of the items (within a particular item type) as independent variables to account for dependencies due to multiple observations per item; inferences are thus holding item constant. To account for multiple observations per examinee, we computed cluster-adjusted standard errors with the clusters being the individual examinees.

Table 3 presents the results for the main coefficient estimates of interest (i.e., those for the ability groups) and for the comparison of each answer change response type (WR, RR, WW) to that of RW. We focus on the comparison between the odds of making RW relative to WR changes

given these represent the beneficial and harmful changes. For SSMC items, the odds of making a WR change instead of RW change is 1.6 times higher for high-ability PSTs than low-ability PSTs holding item constant, whereas the corresponding odds for medium-ability versus low-ability is not significant. Similar results are found for MSMC items. For grid items, the odds of making a WR change instead of a RW change is estimated from the model to be 1.73 times higher for high-ability than low-ability PSTs, but it is not quite significant at the $\alpha = 0.05$ level with a p -value of 0.054. The model results generally show that high-ability PSTs are more likely than low-ability PSTs to make beneficial changes than detrimental changes.

This result is further seen in panel b of Fig. 4, which shows that, for all items and each of the three item types, the proportion of examinees making WR responses increases monotonically as the performance level increases and, for each performance level, the mean proportion making WR changes is higher than RW changes. Accordingly, the item gain-to-loss ratio increases by performance level for each item type (all: 2.6, 2.9, 4.1; SSMC: 2.8, 3.0, 4.8 MSMC: 2.3, 2.7, 3.5; grid: 2.4, 3.4, 3.7). High-performing examinees are about four times more likely to make beneficial than detrimental changes compared to low-performing examinees who are “only” about two times more likely for any item type.

Table 3 Multinomial logistic regression results modeling the odds of each answer change response type (relative to RW) by ability group

Item type	Comparison	Variable		Coefficient estimate	Cluster-adjusted Std. error	Odds ratio	Cluster-adjusted <i>p</i> -value	
SSMC	WR vs RW	Intercept		1.03	0.42		0.014	
		log odds	Ability: Med vs low	0.00	0.13	1.00	0.976	
MSMC	WR vs RW	Intercept		1.49	0.30		0.000	
		log odds	Ability: Med vs low	0.15	0.15	1.16	0.329	
Grid	WR vs RW	Intercept		0.81	0.30		0.008	
		log odds	Ability: Med vs low	0.42	0.24	1.52	0.076	

Fixed effects were also included for each item (i.e., dummy indicators for items), but to streamline results, the estimated coefficients for these variables are not presented. Bold values are significant at the $\alpha=0.05$ level. The intercepts reflect the estimated log-odds of making each response change (e.g., WR, RR, or WW) relative to a RW change for low-ability PSTs for the reference item within each item type. The number of observations used in each of the models by item type varies given there are differing numbers of examinees by item response changes for each item type (Number of person-by-item observations = 3,136, 2,898, and 1,462 for SSMC, MSMC, and grid items, respectively)

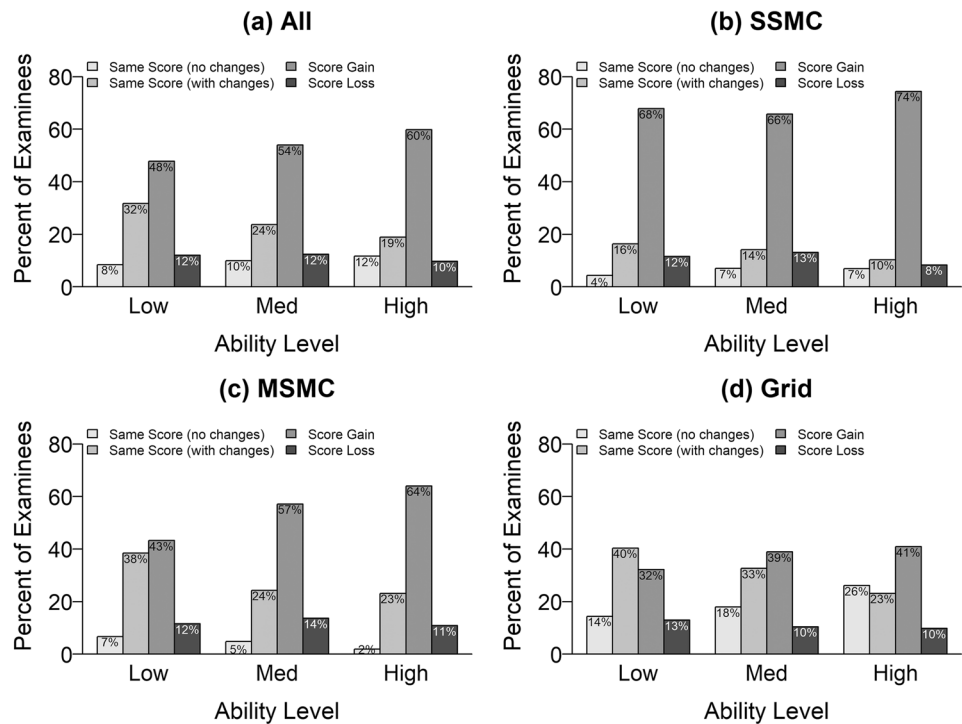
Note. *SSMC* single-selection multiple-choice items, *MSMC* multiple-selection multiple-choice items

Examinee Level

At the examinee level, we are primarily interested in exploring whether high-ability PSTs benefit more from answer changing than low-ability PSTs as has been found in previous studies. First, we review the distribution of total score changes by ability level and item type. Figure 5 shows the distribution of total score statuses for each item type and over all items: It shows the percentage of examinees whose total test scores remain unchanged because they did not make any changes (non-answer changers) and those whose made some changes and then had their total score remain the same, increase, or decrease. For SSMC items, the distributions look similar across ability level and a chi-squared test

of independence between score status and ability level was not significant ($P(\chi^2(6) > 8.72) = 0.19$). For MSMC and grid items, there is a significant relationship between score status and ability level (MSMC: $P(\chi^2(6) > 26.59) = 0.0002$; Grid: $P(\chi^2(6) > 20.83) = 0.002$). Post hoc tests of standardized residuals with Bonferroni corrections for multiple comparisons within each item type reveal which percentages differ significantly from what would be expected if the two variables were independent. Specifically, as evident in Fig. 5, for MSMC items, low-ability PSTs had a higher frequency of answer changers with scores unchanged from answer changing and a smaller frequency of answer changers with score gains than expected if the score status and ability level were independent, whereas high-ability examinees

Fig. 5 The distribution of examinees by total score status change for each ability level and item type



Note. *SSMC* single-selection multiple-choice items; *MSMC* multiple-selection multiple-choice items.

had significantly more score gainers. Thus, for MSMC items, more high-ability PSTs are benefiting from their answer changes than low-ability PSTs. For grid items, low-ability PSTs had significantly more answer changers with scores unchanged, and high-ability PSTs had significantly more non-answer changers (scores unchanged with no answer changes) and significantly fewer answer changers with scores that remained the same after changing than expected if score status and ability level were independent. That is, for grid items, unlike SSMC and MSMC items, the tendency to make any changes was lower for high than low or medium ability PSTs.

Digging deeper, we can examine the extent the mean (total) score change for answer changers varied by ability level. As seen in Table 4, for SSMC and MSMC items, high-ability answer changers gained at least 1.4 times as many score points as low- or medium-ability answer changers. For instance, high-ability answer changers for SSMC items gained 2.07 on average, which is 1.46 times larger than the average gain of 1.42 points for low-ability answer changers. For both item types, overall ANOVAs were significant (SSMC: $P(F(2, 767) > 6.8) = 0.001$; MSMC: $P(F(2, 781) > 10.1) = 0.00005$) and post hoc pairwise comparisons using Tukey's HSD to adjust for multiple comparisons were significant for comparing high to low and high to medium with no significant difference between low and medium ability levels (SSMC: high vs low: $p = 0.004$, high vs med: $p = 0.003$,

med vs low: $p = 0.92$; MSMC: high vs low: $p = 0.00003$, high vs med: $p = 0.006$, med vs low: $p = 0.11$). For grid items, although there were significantly fewer high-ability answer changers, there was no difference in the impact of answer changing among ability levels ($P(F(2, 662) > 2.2) = 0.11$); they all gained about 0.5 points on average.

Discussion

Although there is a depth of literature on answer changing behavior, there is no other study for our integrated construct, examinee population, or set of innovative item types. Largely, our study's findings were consistent with those supported by decades of research; while a minority of item responses were changed, a majority of examinees made answer changes, more answer changers made WR than RW changes, more answer changers reaped test score gains than losses, and higher performing examinees benefited somewhat more than lower performing examinees. These patterns generally held for traditional SSMC items as well as two TEIs—MSMC and grid items—a finding in and of itself.

In some cases, specific findings emerged by item type. For instance, substantially more examinees made WW changes for grid items on average (13.1%) than for other item types (2.6% for SSMC and 7.6% for MSMC). This

Table 4 Overall impact of answer changing on total score by ability level

Statistic	All items (48 items)				SSMC (24 items)				MSMC (17 items)				Grid (7 items)			
	Low	Med	High		Low	Med	High		Low	Med	High		Low	Med	High	
Total examinees	208	411	203		208	411	203		208	411	203		208	411	203	
Number with at least one change	208	410	203		199	382	189		194	391	199		178	337	150	
Percent answer changers	100%	100%	100%		96%	93%	93%		93%	95%	98%		86%	82%	74%	
Mean number items changed for answer changers (SD)	9.45 (4.56)	9.14 (4.74)	8.79 (4.26)		4.19 (2.51)	4.05 (2.48)	3.98 (2.50)		3.68 (2.09)	3.72 (2.12)	3.68 (1.82)		2.35 (1.25)	2.21 (1.11)	1.99 (0.96)	
Mean total score change for answer changers (SD)	2.23 (2.57)	2.60 (2.78)	3.54 (3.12)		1.42 (1.75)	1.49 (1.97)	2.07 (2.21)		0.60 (1.17)	0.84 (1.39)	1.21 (1.48)		0.35 (1.00)	0.49 (0.93)	0.57 (0.95)	

Note. *SSMC* single-selection multiple-choice items, *MSMC* multiple-selection multiple-choice items, *Low* low ability, *Med* medium ability, *High* high ability

result is not surprising given that even if an examinee changed two rows of a grid item to correct, if they had the third row incorrect, they would still have a WW answer change. Due to the higher prevalence of WW changes for grid items, a higher portion of answer changers had scores unchanged for grid items (40%) than for SSMC (15%) and MSMC items (29%). Alternatively, we could have scored the grid items polytomously (instead of dichotomously) where examinees receive a point for each correct row. We re-analyzed our data under this partial credit scoring rule, and although it reduced the percentage of answer changers with unchanged scores from 40 to 22.6% and increased the percentage with score gains from 47 to 53%, it also increased the percentage of answer changers with score losses from 14 to 24% as some of the WW changes under dichotomous scoring would result in a loss of points under partial-credit scoring. Accordingly, the examinee gain-to-loss ratio reduces from 3.4 (= 47/14) to 2.2 (= 53/24). Thus, using a partial credit scoring approach does not necessarily improve the benefit of answer changing for grid items. Further studies could probe the impact of different scoring rules for grid and MSMC items on answer changing benefits and the impact that informing examinees of the scoring rules has on the likelihood that they change their answers. In addition, we observed higher rates of examinees changing responses for the grid and MSMC items on average over the SSMC items. As multipart items that require two or more selections, the TEIs give examinees more opportunities to change responses than SSMC items that require only one selection, which may explain the differential rates. Further research would need to be done to determine if examinees are more likely to change responses to items assessing a rich integrated construct like CKT about matter.

The general patterns of findings were consistent for the TEIs and traditional SSMC items, but the specific magnitudes of outcome statistics often differed across item types. For instance, the examinee gain-to-loss ratio varied from 3.4 for MSMC items to 5.9 for SSMC items, indicating that examinees have a higher chance of benefiting from answer changing on SSMC than MSMC items on our assessment. However, as previously noted, our item types were not intended to be parallel measures with the same item statistics of difficulty and discrimination (see Fig. A1) assessing the same content (see Supplemental Table A1). The differences in item characteristics across item types are of note as they relate somewhat to answer changing behavior. For instance, the percentage of examinees who make a change on an item negatively correlates with item difficulty ($r = -0.53$) and item-total score correlation ($r = -0.28$): Examinees are less likely to make a change to easy, more discriminating items (higher item-total correlations). In addition, the percentage of examinees making beneficial (wrong-to-right) changes correlates negatively with (log) item time ($r = -0.22$),

indicating items with longer response times, on average, have fewer examinees making beneficial changes than for items with shorter response times. A systematic experimental study would be needed to isolate the impact of the item format itself on answer changing from differences in CKT assessed and item difficulty. Additional TEI formats may also be of interest to study.

Our study served as an initial, exploratory step in characterizing answer changing behavior for innovative, online CKT science assessments that assess a construct of importance in the preparation of the teacher workforce. Accordingly, the findings for our instrument have implications for item development, test navigation, and test preparation for such high-stakes licensure tests like *Praxis*®. Examinees operated under their own assumptions about the benefits of answer changing, which may reflect the common belief that it is best to stick with your initial response. Accordingly, we cannot say that encouraging copious answer changing is a good test-taking strategy. However, given that PSTs generally benefited from the answer changes they did make, PSTs should not be overly worried that making a change is a bad strategy. In preparation materials for tests with similar structural designs (e.g., administered online and item revisits are prohibited), PSTs should be encouraged to change their answers regardless of item type if they have misgivings on their initial response. Such behavior could have meaningful impact on a PST's licensure or entry into the workforce. Future studies could investigate how these answer changing behaviors provide insights into how examinees interact with the item content and answer options for specific options, including the distractors. Further research could also examine the relationship with answer changing behavior for such assessments with other examinee characteristics, such as demographic variables or variables pertinent to the population taking the exam. Better understanding about which options examinees are selecting between and the order of particular selections can inform decisions about how to structure items and what information to include or remove from items.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s10956-022-09971-2>.

Funding This study was supported by a grant from the National Science Foundation (Award No. 1813254).

Declarations

Consent Statement The data collection for this study was approved by our institutional review board. All participants completed a written consent form to have their data used for research purposes.

Disclaimer The opinions expressed herein are those of the authors and not the funding agency.

References

- Al-Hamly, M., & Coombe, C. (2005). To change or not to change: Investigating the value of MCQ answer changing for Gulf Arab students. *Language Testing*, 22(4), 509–531.
- Ball, D. L., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education*, 59(5), 389–407. <https://doi.org/10.1177/0022487108324554>
- Bath, J. A. (1967). Answer-changing behavior on objective examinations. *The Journal of Educational Research*, 61(3), 105–107.
- Bauer, D., Kopp, V., & Fischer, M. R. (2007). Answer changing in multiple choice assessment change that answer when in doubt—and spread the word! *BMC Medical Education*, 7(1), 1–5.
- Bertram, A., & Loughran, J. (2012). Science teachers' views on CoRes and Pap-eRs as a framework for articulating and developing pedagogical content knowledge. *Research in Science Education*, 42(6), 1027–1047. <https://doi.org/10.1007/s11165-011-9227-4>
- Bridgeman, B. (2012). A simple answer to a simple question on changing answers. *Journal of Educational Measurement*, 49, 467–468. <https://doi.org/10.1111/j.1745-3984.2012.00189.x>
- Friedman-Erickson, S. (1994). To change or not to change: the multiple choice dilemma.
- Geiger, M. A. (1991a). Changing multiple-choice answers: Do students accurately perceive their performance? *The Journal of Experimental Education*, 59(3), 250–257.
- Geiger, M. A. (1991b). Changing multiple-choice answers: A validation and extension. *College Student Journal*.
- Green, K. (1981). Item-response changes on multiple-choice tests as a function of test anxiety. *The Journal of Experimental Education*, 49(4), 225–228.
- Henze, I., & Van Driel J. H. (2015). Toward a more comprehensive way to capture PCK in its complexity. In A. Berry, P. Friedrichsen, & J. Loughran (Eds.), *Re-examining pedagogical content knowledge in science education* (pp. 120–134). New York: Routledge.
- Jeon, M., De Boeck, P., & van der Linden, W. (2017). Modeling answer change behavior: An application of a generalized item response tree model. *Journal of Educational and Behavioral Statistics*, 42(4), 467–490.
- Liu, O. L., Bridgeman, B., Gu, L., Xu, J., & Kong, N. (2015). Investigation of response changes in the GRE revised general test. *Educational and Psychological Measurement*, 75(6), 1002–1020.
- Mathews, C. O. (1929). Erroneous first impressions on objective tests. *Journal of Educational Psychology*, 20(4), 280.
- McConnell, M. M., Regehr, G., Wood, T. J., & Eva, K. W. (2012). Self-monitoring and its relationship to medical knowledge. *Advances in Health Sciences Education*, 17(3), 311–323.
- Mikeska, J. N., & Castellano, K. (2021, April 9–12). National field test results examining elementary preservice teachers' content knowledge for teaching about matter [Paper presentation]. AERA Annual Meeting, Orlando, FL. Virtual conference.
- National Academies of Sciences, Engineering, and Medicine. (2015). *Science teachers learning: Enhancing opportunities, creating supportive contexts*. Committee on Strengthening Science Education through a Teacher Learning Continuum. Board on Science Education and Teacher Advisory Council, Division of Behavioral and Social Science and Education. Washington, DC: The National Academies Press.
- National Research Council. (2013). *Monitoring progress toward successful K-12 STEM education: A nation advancing?* Washington, DC: The National Academies Press. <https://doi.org/10.17226/13509>
- Ouyang, W., Harik, P., Clauser, B. E., Paniagua, M. A. (2019). Investigation of answer changes on the USMLE Step 2 Clinical Knowledge examination. *BMC Medical Education*, 19(389). <https://doi.org/10.1186/s12909-019-1816-3>

- Park, S., & Oliver, J. S. (2008). Revisiting the conceptualization of pedagogical content knowledge (PCK): PCK as a conceptual tool to understand teachers as professionals. *Research in Science Education, 38*(3), 261–284. <https://doi.org/10.1007/s11165-007-9049-6>
- Park, S., & Suh, J. K. (2015). From portraying to assessing PCK: Drivers, dilemmas, and directions for future research. In A. Berry, P. Friedrichsen, & J. Loughran (Eds.), *Re-examining pedagogical content knowledge in science education* (pp. 104–119). New York: Routledge. <https://doi.org/10.4324/9781315735665>
- Payne, B. D. (1984). The relationship of test anxiety and answer-changing behavior: An analysis by race and sex. *Measurement and Evaluation in Guidance, 16*(4), 205–210.
- Ramsey, P. H., Ramsey, P. P., & Barnes, M. J. (1987). Effects of student confidence and item difficulty on test score gains due to answer changing. *Teaching of Psychology, 14*(4), 206–210.
- Roth, K. J., Garnier, H. E., Chen, C., Lemmens, M., Schwille, K., & Wickler, N. I. (2011). Video-based lesson analysis: Effective science PD for teacher and student learning. *Journal of Research in Science Teaching, 48*(2), 117–148. <https://doi.org/10.1002/tea.20408>
- Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*(2), 4–14. <https://doi.org/10.3102/0013189x015002004>
- Stylianou-Georgiou, A., & Papanastasiou, E. C. (2017). Answer changing in testing situations: The role of metacognition in deciding which answers to review. *Educational Research and Evaluation, 23*(3–4), 102–118.
- Van der Linden, W. J., Jeon, M., & Ferrara, S. (2012). “A paradox in the study of the benefits of test-item review”: Erratum. *Journal of Educational Measurement, 49*(4), 466. <https://doi.org/10.1111/j.1745-3984.2012.00188.x>
- Vidler, D., & Hansen, R. (1980). Answer changing on multiple-choice tests. *The Journal of Experimental Education, 49*(1), 18–20.
- Waddell, D. L., & Blankenship, J. C. (1994). Answer changing: A meta-analysis of the prevalence and patterns. *The Journal of Continuing Education in Nursing, 25*(4), 155–158.
- Wilson, S. M. (2016). Measuring the quantity and quality of the K–12 STEM teacher pipeline (Education White Paper). Menlo Park, CA: SRI International.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.