“What if I don’t know the answer?” Fifth-grade Students’ Responses to Uncertainty in Test-Taking

Ruth A. Childs
Ontario Institute for Studies in Education, University of Toronto

Susan Elgie
Ontario Institute for Studies in Education, University of Toronto

Amanda Brijmohan
Ontario Institute for Studies in Education, University of Toronto

Jinli Yang
Ontario Institute for Studies in Education, University of Toronto

Canadian Journal of Education / Revue canadienne de l’éducation 42:4 (2019)

www.cje-rce.ca

Abstract

Fifth-grade students watched a short video and then responded to multiple-choice items, including several without correct answers. Based on computer-supported stimulated recall and semi-structured interviews, we tested three common assumptions about what students are thinking when they respond to multiple-choice items in spite of being uncertain of the correct answer. We found that none of the assumptions applied to all students. For example, many of the students believed leaving items blank was unacceptable, in part because it might create the impression that they were not trying on a test. Furthermore, although most students recognized when they were uncertain, a few did not.

Keywords: multiple-choice items, response processes, uncertainty, meta-memory, guessing

Résumé

Des étudiants de 5e année ont visionné un court vidéo puis ont répondu à un questionnaire à choix multiple, dont certaines questions ne contenaient pas la bonne réponse parmi les choix. Sur la base de rappels stimulés par ordinateur et d'entrevues semi-structurées, nous avons testé trois présomptions communes concernant ce que pensent les étudiants lorsqu'ils répondent à des questionnaires à choix multiple devant l'incertitude qu'une réponse est bonne. Les résultats démontrent qu'aucune de ces présomptions ne s'applique à tous les étudiants. Par exemple, plusieurs d'entre eux pensaient qu'il était inacceptable de laisser une question sans réponse, en partie parce que cela pouvait créer l'impression qu'ils n'ont pas essayé. Par ailleurs, alors que la plupart des étudiants ont reconnu lorsqu'ils étaient incertains, ce n'est pas le cas de tous.

Mots-clés : questions à choix multiple, processus de réponse, incertitude, métamémoire, deviner

Acknowledgements

Thank you to the fifth-grade students and their parents. Thank you to Amanda K. Ferguson, Marija Glisic, Christine Pouliot, Sara Rattanasithy, Tanya Scerbina, Tian Tang, Mike Tecson, and Jingshun Zhang for their contributions as research assistants on this project. Funding for this study was provided by the Social Sciences and Humanities Research Council of Canada.

Introduction

Imagine someone is answering a series of multiple-choice items in the form of a test. *What* the test-taker responds to each item is easy to study: We can look at the answer sheet or, if the test was administered on the computer, an electronic file. If the item has four response options labelled A to D, the response record will contain a series of those letters and, possibly, blanks where the test-taker did not respond.

How the test-taker responds is more difficult to study. They may have read the item and selected the response option they believed was correct. They may have guessed, with or without first reading and trying to answer the item. If guessing, they may have chosen randomly among all of the response options, or eliminated some options as unlikely before choosing among those that remained. They may have chosen to leave the item blank. To complicate matters further, they may have used different approaches for different parts of the test.

Although difficult to study, how test-takers respond is important. As the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014) emphasizes, "some construct interpretations involve more or less explicit assumptions about the cognitive processes engaged in by test takers" (p. 15)—that is, how we interpret test-takers' responses often depends on what we assume about their *response processes*. We may assume, for example, that test-takers are conscious of scoring rules, such as penalties for incorrect answers that are intended to discourage guessing, and that their responses are affected by these rules. We may assume that test-takers will leave an item blank if they are uncertain of the correct answer (and that they are able to judge their level of certainty accurately), or we may assume that they do not consider it acceptable to leave a blank. To ensure that we are not misinterpreting test-takers' responses, we need to have a better understanding of which assumptions are likely to hold for whom, under what conditions, on which types of items.

In this study, we investigate response processes of fifth-grade students when they are uncertain of the correct answers to multiple-choice items. Using computer-supported stimulated recall and semi-structured interviews, we test common assumptions about how students respond to uncertainty when answering multiple-choice items.

Perceptions of Uncertainty

The *Oxford English Dictionary* defines uncertainty as "the state or character of being uncertain in mind; a state of doubt; want of assurance or confidence; hesitation; irresolution" ("Uncertainty," definition 3a, 1989). It is possible to have different degrees of uncertainty, to be more or less uncertain (or more or less certain).

In Nelson and Narens's (1990) metamemory framework, judgements related to uncertainty are aspects of *monitoring* that inform *control* processes. Relevant to test-taking, in monitoring, individuals make *feeling-of-knowing* judgements about whether something they cannot recall at the moment is something they do know and might be able to recall in the future, and *confidence* judgements about whether something they have recalled is correct. These are both examples of monitoring. The individual does not have to be conscious of their feeling-of-knowing and confidence judgements for those judgements to affect both what they are conscious of and what they choose to convey to others (Efklides, 2008). In a testing context, confidence judgements may help individuals control whether or not they respond to an item, although, as Koriat and Goldsmith (1996) demonstrated in a series of experiments, *situational demands*, such as incentives for accuracy, may also affect decisions of whether to respond.

There is a large body of research on how often test-takers decide not to respond to an item (see Köhler, Pohl, & Carstensen, 2017, for a recent summary). There is also a long history of research on situational demands in testing—in particular, the effects of scoring rules and of the wording of test instructions (e.g., Bauer, 1971; Bereby-Meyer, Meyer, & Budescu, 2003; Lord, 1975; Prieto & Delgado, 1999; Swineford & Miller, 1953). For example, in a paper titled "Who Is Penalized by the Penalty for Guessing?" Sherriffs and Boomer (1954) wrote:

> We believe that when the right-minus-wrong instructions are added to an ordinary examination a new set of decisions faces the individual student. For example, such questions must necessarily arise as: "How certain am I of this answer? What is the probability of my being right? If the probability is 80 to 20 that I am right then am I better off to 'guess' or to omit the item?" Ability to "figure the odds," subjective levels of confidence in [one's] own judgment, and willingness to

> gamble or to take a chance, are among the variables which might well be expected to influence the right-minus-wrong measure of achievement. (p. 2)

As Sherriffs and Boomer (1954) suggest, confidence judgements are not always accurate. How accurate test-takers are in these judgements—how well *calibrated* they are—has been studied using a variety of scales. For example, Fischhoff, Slovic, and Lichtenstein (1977) asked test-takers to report the probability (out of 1) or the odds (as a ratio) that each of their responses was correct. A consistent finding of such studies (e.g., Kasperski & Katzir, 2013) is that most people are overly confident that their responses are correct, but that there is considerable variability both across individuals and, within individuals, across items.

On what do test-takers base their confidence judgements? Dinsmore and Parkinson (2013) asked students at a university in the United States to rate their confidence in their responses to multiple-choice items, either by marking a point on a line labelled "not confident" at one end and "very confident" at the other, or by comparing their relative level of confidence between items. For each item, they also asked the students to write an explanation of "how you arrived at or what you considered when making your confidence judgment" (p. 8). They found that students referred in their explanations to their prior knowledge about the topic, to information in the text on which the items were based, and to specific wording in the items. A few also mentioned guessing.

Studying the Effects of Uncertainty

Investigating test-takers' response processes when they are uncertain is difficult for two reasons: (1) performance on tests of academic knowledge and skills is affected by the different opportunities test-takers have had to learn the materials, and (2) some test-takers may not be uncertain about any items. In a series of studies with elementary students in Switzerland, Roebers and her colleagues addressed the first difficulty by presenting novel information—for example, a video about the production of sugar (Krebs & Roebers, 2010; Roebers, 2006) or pictures of Japanese characters (Roderer & Roebers, 2010)—to the students before testing their memory of the information, ensuring that all students had the same opportunity to learn the materials.

To address the second difficulty, some researchers have asked unanswerable questions—that is, “questions that ask about information that the child has never encoded or stored in memory, and are therefore not linked to the typical processes of retrieval” (Waterman & Blades, 2013, p. 215). For example, in an effort to understand children’s reliability as witnesses in court proceedings, Waterman and Blades (2013) had children watch a staged event and then, after a delay, asked them questions about it, including some unanswerable questions. The length of the delay and the children’s verbal ability predicted whether they accurately reported that they did not know the answer to the unanswerable questions. In their research, Roebers and her colleagues have also included unanswerable items, so that all test-takers, regardless of knowledge or ability, should be uncertain of their response to those items. For example, Roderer and Roebers (2010) taught 7- and 9-year-old children the meanings of several Japanese characters (none of the children had previously seen the characters) and then tested their recall. The recall test included a few new characters. Both the 7- and 9-year-old children gave lower confidence judgements to their responses to the new characters compared to those for the characters they had studied, though the difference was larger for the 9-year-old children. To see if the children might have considered a different confidence rating than they ultimately chose, the authors also analyzed the amount of time the children looked at each rating on the confidence scale; the children looked longest at lower ratings for the new characters. Interestingly, only a few of the children said they noticed that some of the characters were new, and that was only after being debriefed about the study. Roebers (2006) and Krebs and Roebers (2010) have studied children’s confidence judgements on other types of recall tests and have consistently found that older children are a little more accurate in their judgements of uncertainty when faced with unanswerable items (that is, they rate their confidence lower or are more likely to choose a “don’t know” option). Krebs and Roebers (2012) found that higher-achieving children were more accurate in their judgements of uncertainty.

One of the things that is striking in studies of test-takers’ confidence is how rarely uncertainty leads to nonresponse, even to items for which the test-takers cannot possibly know the correct response. For example, even when, as in one of the conditions of Krebs and Roebers’s (2010) study, the researchers told test-takers that they would earn one point for each correct response, but lose three points for each incorrect response, and then instructed them to cross out any answers they thought might be incorrect, 8- to 11-year-olds kept responses (automatically counted as incorrect) to half or more of the unanswerable items.

One possibility to explain this pattern of responses is that the conditions of the study or test led test-takers to believe they should respond, even if uncertain. Test-takers may, quite sensibly, assume that tests will only include items that give them opportunities to demonstrate their knowledge. Schwarz (1994) has critiqued the related practice of including unanswerable survey questions in survey design experiments and then concluding, when study participants provide an answer, that they are afraid of seeming to be uninformed and so "generate some random response, apparently confirming social scientists' wildest nightmares" (p. 135). He argues that the participants are making a rational response to a very unusual situation: "What is at the heart of reported opinions about fictitious issues is not that respondents are willing to give subjectively meaningless answers, but that researchers violate conversational rules by asking meaningless questions in a context that suggests otherwise" (p. 136).

Another possible reason students—especially young children—might respond to items intended to be unanswerable, was explored by Hughes and Grieve (1980). In their study, 5-year-old and 7-year-old children were asked questions intended to be unanswerable, such as "One day there were two people standing at a bus-stop. When the bus came along, who got on first?" All of the children in the study provided answers to this question, often by referring to norms of queuing, such as "the one there first," or by providing additional details that would make it answerable, leading Hughes and Grieve to suggest that researchers should be mindful that children are not "passive recipient[s] of questions and instructions," but are "actively trying to make sense of the situation" (p. 160).

If test-takers are reluctant to leave an item blank, perhaps their uncertainty may be seen in how often they change their responses. In a study of university students taking a final exam, Stylianou-Georgiou and Papanastasiou (2017) recorded the changes students made and also asked them to rate their confidence. For the items that were changed by the largest number of students, they examined the reasons students selected for making changes. Consistent with the changes being related to uncertainty, the most frequently endorsed reason was "Rethought and reconceptualized the answer."

Studying Response Processes

Asking test-takers, during testing, to verbalize their thinking may provide some insight into how and why test-takers respond the way they do (Leighton, 2004). However, it is possible that being asked to think aloud while responding alters how test-takers respond to items—for example, they might assume they must respond to every item because the researcher is monitoring their response. This has led some researchers to ask test-takers to report their thinking *after* they have finished the test. For example, Jakwerth, Stancavage, and Reed (2003) interviewed eighth-grade students immediately after they took a National Assessment of Educational Progress (NAEP) test, asking them to explain their decisions not to respond. Lack of knowledge and lack of time were the two most often cited reasons. Jakwerth et al. found that 46% of those test-takers who omitted an item (i.e., did not respond to an item but responded to subsequent items) reported understanding the item without knowing the answer, while 51% either did not understand the item as a whole or did not understand one or more of the words. Some of these test-takers reported that they had planned to return to the omitted items, but ran out of time. Because Jakwerth et al. were particularly interested in nonresponse and interviewed test-takers after they had finished taking the test, they were able to review the test booklets, identify those test-takers who had left some items blank, and then select them to be interviewed.

Jakwerth et al.’s (2003) study is an example of *stimulated recall*, in which “some tangible (perhaps visual or aural) reminder of the event [stimulates] recall of the mental processes in operation during the event itself” (Gass & Mackey, 2016, p. 14). As Gass and Mackey (2016) explain, stimulated recall

> has an advantage over a simple post hoc interview in that the latter relies heavily on memory without any prompts and it has an advantage over think-aloud protocols in that for think-alouds the researcher needs to train participants, and even after training, not all participants are capable of carrying out a task and simultaneously talking about the task. (pp. 15–16)

Collecting verbal reports after test-takers answer items on a computer is especially promising for research in this area, as the computer can capture response times and can record when test-takers change responses or skip and return to items (e.g., Douglas & Hegelheimer, 2007).

This Study

In this study, we investigated how fifth-grade students respond when they are uncertain of the correct answer to a multiple-choice item. We sought to control for differences in students' prior knowledge by asking the students to respond to multiple-choice items that measured recall of information from a short video. We sought to ensure students would be uncertain by including several items that were unanswerable either because the information was not provided in the video or because the set of response options did not include the correct response. We did the study in Ontario, and recruited fifth-grade students in the belief that they were old enough to explain their thinking. All the students in the study had participated in a province-wide large-scale assessment that included multiple-choice items when they were in the third grade.

Focusing on the students' responses to the unanswerable items, we wanted to know: (1) What do the students report that they usually do when uncertain about an answer? (2) What did they do on this test when uncertain about an answer? and (3) Are the following common assumptions about how students respond to uncertainty when answering multiple-choice items consistent with these data: (a) that students realize when they are uncertain, (b) that they make independent decisions for each item about whether to guess or leave the item blank, and (c) that, when they guess, students are motivated by the desire to "maximize" their score?

Methods

Participants

After receiving approval from the University of Toronto's Research Ethics Board, we recruited fifth-grade students from the greater Toronto and London, Ontario, areas through posters in local community centres and emails to several communities of parents. Interested parents were directed to a website to schedule an appointment for their child. Most parents brought their children to a computer lab on the university campus, although six students participated in their homes. Students received a $30 bookstore gift card.

Six students participated in a pilot test of the computer interface and the stimulated recall and interview protocols in February 2014. After the pilot test, the interface and protocols were revised and 34 students participated in March 2014. After removing six students who did not see the first three items because of an ambiguity in the interface, 28 students (16 females and 12 males) were included in the analyses. All but one was attending a publicly funded school. All were fluent in English. Eleven spoke or were learning a language other than English and French from their parents (e.g., Arabic, Croatian, Ewe, Hindi, Italian, Japanese, Konkani, Mandarin, Spanish). Parents reported that the children's report card grades ranged from As to Cs.

Two of the participants had attended third grade in another province, but the rest can be assumed to have taken the provincial assessment, which included multiple-choice items and the following instructions: "Be sure to attempt all questions. If you leave a question blank, it will be scored 0."

Procedure

Using a program written for this study in Flash and running on a laptop computer, the students were shown a short video about how honey is produced and then answered multiple-choice items such as:

4. How do the bees get the water content of the nectar to evaporate?
- They fan it. *correct response
- They dance on it.
- They suck it out.
- They chew it.

Five of the 20 items were not answerable from information in the video; three items were covered in the video, but did not include a correct option; one was not covered in the video; one was not covered in the video and also had nonsensical options. There was no time limit for the students to answer the items. As the students answered the items on a laptop computer, screen capture software recorded the computer screen, including the movements of the cursor. Using a stimulated recall approach, after the student finished responding to all 20 items, the researcher administering the task viewed the screen recording with the student, asking them to describe what they had been thinking as they were moving the cursor and selecting responses on the computer screen. All items, whether answerable or not, were included in the stimulated recall. The researcher also interviewed the students, asking about their test-taking strategies, previous test-taking experience, and attitudes toward tests. The full session was audio recorded and lasted 30 to 50 minutes.

Analyses

The time to respond to each item, the response(s) selected, and the sequence of responses were captured in a computer log by the software and were subsequently summarized in a spreadsheet. Graphs were created to represent each student's movement through the items. Descriptive analyses were performed on the response patterns.

The stimulated recall responses and interviews were transcribed by the research team. The stimulated recall transcripts were synchronized with the screen recordings in NVivo. Each transcript was coded independently by at least two members of the research team for reported past responses to uncertainty, reasons for these responses, explanations for responses to the unanswerable items, and reasons for changing answers or skipping items.

Results

As Table 1 shows, the students varied in their performance, both in terms of the number of answerable items answered correctly (ranging from six to 14 out of 15 answerable items) and the amount of time they took to respond to all 20 items (ranging from 1.1 to 7.6 minutes). The correlation between the number of correct answers and total time spent was only .05 and not statistically significant (p = .80). (Note: The unanswerable items are presented in italics in the text and tables.)

Table 1. Items not answered and skipped by each student (ordered by number correct and, within number correct, by time)

Student	Items skipped	Answers changed	Number correct	Time (minutes)
S16		Item 9 (C,D*); Item 16 (A,B,A,D*); *Item 18(A,B)*	14	6.52
S08	*Item 15 (not answered)*	Item 16 (C,A,D*)	14	5.02
S12		Item 9 (C,D*); Item 19 (B,C*); *Item 20 (B,A)*	14	2.93
S19			13	4.82
S23	*Answered at end: Item 11, Item 15*	Item 13 (A,B*); Item 16 (A,C,D*)	13	4.82
S09		Item 14 (B,C*)	13	4.72
S30			13	4.52
S37		Item 13 (A,B*)	13	4.23
S15	*Item 11 (answered after Item 16)*	Item 16 (A,B,A)	13	4.13
S25	Item 19 (answered at end)	Item 1 (A,C*); *Item 11 (D,A)*; Item 14 (A,C*); *Item 15 (C,D)*; Item 16 (A,D*); Item 17 (D,C)	13	3.50
S24		Item 1 (A,C*); Item 9 (C,D*,C); Item 13 (A,B*); Item 16 (A,D*)	13	2.98
S11		Item 1 (A,C*); Item 3 (D*,A); Item 4 (C,A*); Item 5 (B,C*); *Item 11 (B,A); Item 20 (B,C,B,C)*	12	5.83
S17		*Item 15 (C,D,C)*	12	4.08
S07			12	4.07
S21	Answered at end: Item 5, Item 9, *Item 11*, *Item 15*	Item 13 (A,B*,C); *Item 20 (A,B,A)*	11	7.62
S39		Item 16 (A,C,A,C,A,C,A,C,A,C,A,D*)	11	6.27
S27		*Item 15 (C,D,C); Item 20 (B,A)*	11	6.05
S36	Item 9 (not answered)	Item 5 (A,B); Item 19 (D,C*)	11	5.62
S31		Item 9 (B,C)	11	3.95
S32		Item 4 (A*,D); Item 5 (A,C*)	11	1.12
S29		Item 3 (C,A,C); Item 10 (A,B*); Item 13 (A,B*); Item 16 (A,D*); *Item 20 (A,B)*	10	4.93
S13	*Item 11 (answered at end); Item 15 (answered after Item 16)*	Item 13 (C,D); Item 16 (A,D*)	10	4.32

Student	Items skipped	Answers changed	Number correct	Time (minutes)
S14		Item 4 (C,D); Item 16 (A,D*); *Item 18 (C,A)*; Item 19 (C*,B,A)	9	6.40
S20		*Item 15 (C,B,C)*; Item 19 (D,B)	9	5.95
S18		Item 9 (B,A)	9	4.63
S26		Item 9 (D*,B); Item 10 (A,C); *Item 20 (A,B,A)*	7	3.58
S35		Item 12(C,B*)	7	3.45
S10	*Item 7(not answered)*	Item 2 (C,A,D); Item 14 (C,D); *Item 15 (C,B)*; Item 17 (D,C)	6	3.40

Note. In the second and third columns, items shown in italics were unanswerable. In the third column, an asterisk indicates a correct response. The test contained 20 items in total; five had no correct response.

What Do the Students Report that They Usually Do When Uncertain About an Answer?

In the interview that followed the stimulated recall of what they were thinking while responding to individual items, the students were asked how they respond when uncertain of the answer to an item: by guessing or by leaving the item blank. They were asked if they ever skip an item (leave it blank temporarily) with the intention of returning and answering it later. For guessing, leaving blank, and skipping-and-returning, they were also asked if each was ever okay to do and why or why not.

All 28 of the students said they sometimes skip and return to an item. Of the 16 students who gave reasons for using this strategy, eight mentioned that sometimes subsequent items will provide a clue to the answer of the skipped item (e.g., “Sometimes you can get little hints in bits and pieces in questions after” [S16]) and nine described the danger of spending too much time on an item and running out of time for other items (e.g., “You don’t want to waste all of your time on one question, just looking at the question, because you could’ve answered other questions” [S19]).

Most of the students (23 out of 28) said they sometimes guess and that it is never okay to leave an item blank. The exceptions were four students who said that it is okay to leave an item blank, although they would sometimes guess (one of the four specified that it is okay to leave an item blank only when the item is unanswerable: “If you know that it’s not one of the answers then it’s fine to leave it blank” [S14]), and one student who would neither guess nor leave an item blank because "I always have to try my best on whatever test I get" [S31]).

As reasons for guessing, 18 cited the possibility that the guess might be correct and so could increase one's score on the test (e.g., "You could get extra marks that way… you never know if it could be right" [S10]). Eight students worried that leaving an item blank could suggest they were not giving sufficient effort (e.g., "At least you're thinking about it and the teacher might be able to see that" [S32]).

When asked if they remembered any test-taking advice their teachers gave them when they were taking the provincial assessment in third grade, only four mentioned the importance of answering all items (e.g., "Do not leave any questions blank" [S29]).

What Did They Do on This Test When Uncertain About an Answer?

Awareness of unanswerable items. We wondered whether the students would suspect that some of the items were intended to be unanswerable. A few were explicit in their interviews: "I thought, 'this is rigged'" (S25); "I wondered if they were typos" (S15); "I stopped and thought to myself that I don't know if I misheard that or if it doesn't have the answer" (S36). Many of the students stated in the stimulated reviews that they recognized that the correct answers were not present for the two unanswerable items that were covered in the video but did not include the correct response among the options. The first such item was *Item 7*:

> 7. What does the beekeeper do to warn bees that foreigners are about to enter the hive?
>
> - Knocks on the beehive
> - Plays music
> - Heats the beehive
> - Shakes the beehive

The video stated: "The beekeeper sprays the hive with smoke from burning pine needles, a scented warning that foreigners are about to enter the hive." This was not among the response options, however. Nevertheless, 19 of the 28 students recalled the correct response and noted that it was not included in the item.

Similarly, *Item 11* asked, “What does the bee escape smell like?” The correct response was “cherries,” but that was not among the response options. Twenty-one of the students recalled the correct response and recognized that it was absent. All of the students provided answers, however. They offered explanations of their process of selecting a response:

- *Strawberries* sounds like *cherries* (e.g., “Well I think they said cherries, but the end of strawberries and cherries sound alike and nothing else sounds like cherries in there so I picked strawberries.” [S37])
- *Chocolate* starts with “ch,” like *cherries* (e.g., “Because those both [“cherries” and “chocolate”] start with a ‘c’ and they go ‘ch’ and ‘ch’ so I was thinking, maybe I just heard it wrong.” [S29])
- Strawberry is a sort of berry (e.g., “I remembered they said something about a cherry smell. But that wasn’t one of the choices so I picked a different berry—one that’s most related to it.” [S30])
- Licorice can be cherry-flavoured (e.g., “I picked licorice because [the video] said it smelled like cherries, and licorice is cherry flavoured, there’s some licorice that’s cherry flavoured, so I went with licorice.” [S16])

The first and last have to do with the sounds of the words. The second is an example of the students generalizing to a higher category: berries. The third involves students relating the information in the video to previous knowledge.

The other three unanswerable items were not addressed in the video. *Item 15* asked, “What company makes the honey extractor?” No company name was mentioned in the video nor was one visible on the machines; the response options offered were a mix of real company names (Honeywell, best known for its heating and cooling systems, not honey-processing machinery) and names created to sound as though they might relate to the honey extraction function (Combvac) or to bees (Beesley and Beesum). One student mentioned that “the names were puns” (S25); three other students recognized that the answer was not provided in the video.

Item 18 asked, “What is the boiling point for honey?” The video mentioned that granulated honey turns back into a liquid at 130 degrees Fahrenheit, but did not mention honey’s boiling point (furthermore, the response options for *Item 18* ranged from 200 to 230 degrees Fahrenheit). One student correctly reported that the answer was not in the video. This question may have been particularly confusing for these students, as they are more familiar with temperatures being expressed in degrees Celsius than in degrees Fahrenheit.

Finally, *Item 20* was intended to be recognizably absurd:

20. How much honey does each bee need to eat per day?

- Up to 1 pound
- Up to 1.5 pounds
- Up to 2 pounds
- Up to 2.5 pounds

In the stimulated recalls, three students noted that the item was absurd.

Although the unanswerable items were intended to ensure that students were uncertain, nine students reported being certain of one or more of their responses to the unanswerable items. The number varied by item, from two reporting certainty on *Item 15*, to five students reporting certainty on *Item 18*.

Guess. A few students reported guessing on some of the answerable items, as well as the unanswerable ones. However, when the researcher administering the task probed further, the students described narrowing the possible responses to two or three of the four options, based on information from the video or their background knowledge. For example, for Item 3, “The bees break the complex sugar down into which two simple sugars?” Student S17 initially said, “I didn’t know about this, but I took a random guess.” However, when asked “Was there anything about that that came to your memory?” the student reported remembering part of the information: “I remembered one was glucose, but I didn’t remember any of the other ones.”

Leave the item blank. Were students more likely to leave blank the items on which they could not be certain of the correct response—the unanswerable items? As Table 1 shows, only three of the 28 students left any items blank (and only two left unanswerable items blank). That is, 26 of 28 students answered *all five* of the unanswerable items and the remaining two answered *four of the five* unanswerable items.

Skip and return. If most students did not leave items blank, were there any other effects of uncertainty observable in the response patterns? Because each student's movement through the items was recorded, we were able to analyze when they skipped an item but later returned to it (i.e., left the item temporarily blank). We observed two distinct types of skipping: (1) skipping, but returning within a few items (in Figure 1, see Student S13's response to *Item 15* as an example), and (2) skipping, but returning at the end of the test (in Figure 1, see Student S13's response to *Item 11*).

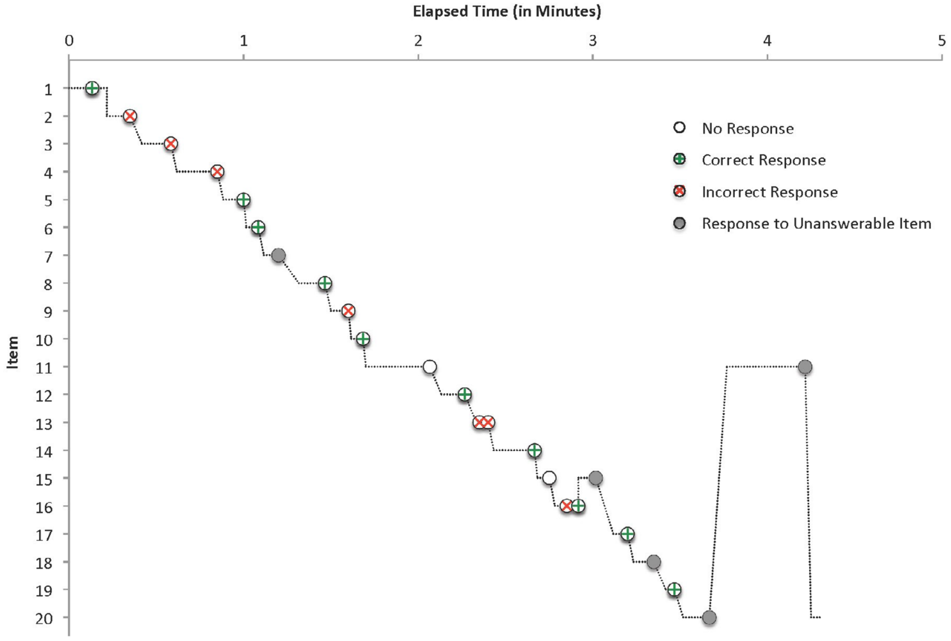

Figure 1. Student S13's response pattern

As summarized in Table 1, students were more likely to skip and return to the unanswerable items. Of the five students who skipped items, three did so only on unanswerable items, one skipped two unanswerable items and two answerable items, and one skipped only one answerable item.

As Table 2 shows, skipping-and-returning and leaving an item blank were also related to item difficulty as measured by the total number of students in this sample who answered the question correctly. The three answerable items that were skipped or left blank were among the four most difficult items. Not surprisingly, most of the items that were skipped or left blank were unanswerable. The number of items skipped or left blank was not significantly related to the student's score (number correct of the 15 answerable items; $r = .01$, $p = .95$).

Table 2. Response changes by item

	Number of students changing responses						Difficulty
Item	Incorrect to correct	Correct to incorrect	Incorrect to incorrect	Unanswerable	Not Answered	Skip and Return	(Proportion of students responding correctly)
Answerable Items							
Item 12	1						1.00
Item 6							.93
Item 8							.93
Item 10	1		1				.89
Item 14	2	1					.86
Item 16	9		1				.86
Item 1	3						.82
Item 2			1				.79
Item 13	3		2				.79
Item 4	1	1	1				.64
Item 17			2				.61
Item 19	2	1	1			1	.61
Item 3		1	1				.57
Item 5	2		1			1	.54
Item 9	2	1	3		1	1	.29
Unanswerable Items							
Item 7					1		
Item 11				2		4	
Item 15				5	1	3	
Item 18				2			
Item 20				6			

Change their response. Another possibility is that students might be more likely to change their responses to items on which they could not be certain. However, as Table 2 shows, the number of students changing responses ranged from zero to 10 for the answerable items and from zero to six for the unanswerable items. The average number of students changing responses was 3.0 for both answerable and unanswerable items.

Although students were not more likely to change their responses to the unanswerable items, it is interesting to consider their changes to the answerable items. As shown in Table 2, for the answerable items (when students changed a response multiple times, only the first and last responses are counted in Table 2), students more often changed their responses from incorrect to correct (26 times) than from correct to incorrect (five times) or between incorrect options (14 times). The correlation between an item's difficulty and the number of students changing responses to an item is a modest -.31 (*p* = .26).

All but three of the students changed at least one answer. Two students changed answers to six of the 20 items. The correlation of the number of items changed with the number correct items was -.14 (*p* = .47). Interestingly, students who changed answers did not take longer on the test: The correlation of the number of changes with the time to answer all the items was .01 (*p* = .98).

In the stimulated recalls, students described five reasons for changing answers:

- Not reading all the response options before answering (e.g., "[On Item 16], I put candles first because I only saw candles and then went down [to the other response options]. I remember candles and lipstick, so, it has to be 'all of the above' because I can't do both." (S13); "[On Item 9], I switched it because I think it said hundreds of thousands of bees, but I hadn't fully scrolled down." [S16])
- Continuing to reason from pre-existing knowledge (e.g., "[On Item 4], I knew it wasn't 'they danced on it' because I don't really think bees can dance. And then I wasn't really sure; I was just thinking about maybe they fanned it? But how would they fan it?" [S11])
- Incompatibility with other items (e.g., "[On *Item 20*], I thought it was that one, and then I went, I looked back up and I saw—I thought it makes about 7 pounds per day, so I think it would be just 1 pound." [S27])
- Discomfort with a previous guess (e.g., "[On *Item 20*], a lot of times it's like, 'that's not the right one, I'm going to choose another one.'" [S11])
- Remembering additional information from the video (e.g., "[On Item 14], I saw 'shakes' and then I went to the next one and I'm like, 'hold on! [It] doesn't shake, it goes *zzzt.*'" [S25])

Discussion

We set out to answer three research questions in the analyses: (1) What do the students report that they usually do when uncertain about an answer? (2) What did they do on this test when uncertain about an answer? and (3) Are the following common assumptions about how students respond to uncertainty when answering multiple-choice items consistent with these data: (a) that students realize when they are uncertain, (b) that they make independent decisions for each item about whether to guess or leave the item blank, and (c) that, when they guess, students are motivated by the desire to "maximize" their score?

The answers to our first two questions were clear. Students reported it is important to answer every question on a test, even when they were not certain of the answer. On our test, they sometimes guessed, sometimes skipped and returned to or changed their answers, but rarely left an item blank. When asked, students responded that these were their usual response habits. These results are consistent with previous research.

In answer to our third question, whether common assumptions of test developers, administrators and users concerning student uncertainty were consistent with the data, we found that none of the assumptions applied to all students. The first assumption was that students are aware of when they are uncertain. We found that, although most of the students recognized the unanswerable items, a few did not and were certain of their responses to one or more of those items. The second assumption was that students make independent decisions for each item about whether to guess or leave the item blank. In fact, our results suggest that students have a strong response set that applies to the entire set of items; at the item level they may be deciding only what to guess, not whether to guess. The third assumption was that, when they guess, students are motivated by the desire to "maximize" their score. We found that some students were more concerned that leaving an item blank might give their teacher the impression that they were not trying.

Returning to Nelson and Narens's (1990) metamemory framework, we would conclude that the students who were certain of their response to unanswerable items judged their *confidence* inaccurately. Many accurately reported that they did not know the answers to at least some of the unanswerable items—that is, their *feeling-of-knowing* was accurate. However, even when their monitoring was accurate, it did not lead them to withhold the response, which would have been an example of control. Because it seems that many of the students did not consider withholding a response to be an acceptable option, they were unable to exhibit control in this way.

As predicted by Koriat and Goldsmith (1996), the students' awareness of *situational demands* seems to have affected their decisions about whether to respond, with some saying they would not leave items blank because they might increase their score if they answered correctly (although we did not discuss whether or how the test would be "scored," several students explicitly stated that they assumed they would not be penalized if they guessed incorrectly, so that they could only gain by guessing) or because it might suggest to their teacher (or presumably the adult researchers in this study) that they were not trying.

Limitations

This study involved a convenience sample of a single-age group of students in a single jurisdiction. It was a low-stakes test, as students were well aware. The data were collected most often in a university computer lab in the company of one or two researchers, a situation unlike usual test administration conditions. Certainly, knowing that they would have to talk about their responses may have affected how the students responded, as might have our presence. Consequently, the generalizability of the results to other testing situations may be limited.

Conclusion

Students had a range of responses to the ambiguity incorporated in the parts of the test. Common to most students was the belief that all test items should be answered. Although students varied in their patterns of response—for example, some answered the items in order and never looked back, while others made many changes to their answers—these variations were unrelated to both achievement and elapsed time. Reactions of some students to the uncertainty intrinsic to the test were highly creative; most students went to considerable effort to provide an answer to our questions. Some students were quite aware of the uncertainty of their responses; others not at all. The results suggest that the assumptions of teachers, administrators, and test developers about student response processes may not reflect students' complex patterns of cognition.

If our assumptions about response processes are inaccurate, that will have implications for how we interpret test responses. For example, in this study, some students' belief that all items should be answered resulted in unexpected results, introducing variance unrelated to the intended construct. When giving students advice about test taking, teachers and test administrators may want to consider carefully how some students may interpret the advice and how that may affect the validity of the use of the test results. Teachers may also look for opportunities to help children become more aware of when they are uncertain and to develop strategies for dealing with uncertainty. In general, this study demonstrated that, instead of assuming that all students are using the same processes to decide whether and how to respond to test items—and that we know what those processes are—we can learn much by asking the students themselves.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Bauer, D. H. (1971). The effect of test instructions, test anxiety, defensiveness, and confidence in judgment on guessing behavior in multiple-choice test situations. *Psychology in the Schools*, *8*(3), 209–215. https://doi.org/10.1002/1520-6807(197107)8:3<208::AID-PITS2310080303>3.0.CO;2-B

Bereby-Meyer, Y., Meyer, J., & Budescu, D. V. (2003). Decision making under internal uncertainty: The case of multiple-choice tests with different scoring rules. *Acta Psychologica*, *112*, 207–220. https://doi.org/10.1016/S0001-6918(02)00085-9

Dinsmore, D. L., & Parkinson, M. M. (2013). What are confidence judgments made of? Students' explanations for their confidence ratings and what that means for calibration. *Learning and Instruction*, *24*, 4–14. https://doi.org/10.1016/j.learninstruc.2012.06.001

Douglas, D., & Hegelheimer, V. (2007). *Strategies and use of knowledge in performing new TOEFL listening tasks* (Draft Final Report to Educational Testing Service). Ames, IA: Iowa State University.

Efklides, A. (2008). Metacognition: Defining its facets and levels of functioning in relation to self-regulation and co-regulation. *European Psychologist*, *13*, 277–287. https://doi.org/10.1027/1016-9040.13.4.277

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance*, *3*, 552–564. https://doi.org/10.1037//0096-1523.3.4.552

Gass, S. M., & Mackey, A. (2016). *Stimulated recall methodology in applied linguistics and L2 research* (2nd ed.). New York, NY: Routledge.

Hughes, M., & Grieve, R. (1980). On asking children bizarre questions. *First Language*, *1*(2), 149–160. https://doi.org/10.1177/014272378000100205

Jakwerth, P. M., Stancavage, F. B., & Reed, E. D. (2003). *An investigation of why students do not respond to questions* (National Center for Education Statistics, Working Paper No. 2003-12). Washington, DC: US Department of Education. https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=200312

Kasperski, R., & Katzir, T. (2013). Are confidence ratings test- or trait-driven? Individual differences among high, average, and low comprehenders in fourth grade. *Reading Psychology*, *34*, 59–84. https://doi.org/10.1080/02702711.2011.580042

Köhler, C., Pohl, S., & Carstensen, C. H. (2017). Dealing with item nonresponse in large-scale cognitive assessments: The impact of missing data methods on estimated explanatory relationships. *Journal of Educational Measurement*, *54*(4), 397–419. https://doi.org/10.1111/jedm.12154

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review*, *103*(3), 490–517. https://doi.org/10.1037/0033-295X.103.3.490

Krebs, S. S., & Roebers, C. M. (2010). Children's strategic regulation, metacognitive monitoring, and control processes during test taking. *British Journal of Educational Psychology*, *80*, 325–340. https://doi.org/10.1348/000709910X485719

Krebs, S. S., & Roebers, C. M. (2012). The impact of retrieval processes, age, general achievement level, and test scoring scheme for children's metacognitive monitoring and controlling. *Metacognition and Learning*, *7*, 75–90. https://doi.org/10.1007/s11409-011-9079-3

Leighton, J. P. (2004). Avoiding misconception, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice*, *23*(4), 6–15. https://doi.org/10.1111/j.1745-3992.2004.tb00164.x

Lord, F. M. (1975). Formula scoring and number-right scoring. *Journal of Educational Measurement*, *12*(1), 7–11. http://dx.doi.org/10.1111/j.1745-3984.1975.tb01003.x

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (pp. 125–173). San Diego, CA: Academic Press.

NVivo 10 [Computer software]. Doncaster, Australia: QSR International.

Prieto, G., & Delgado, A. R. (1999). The effect of instructions on multiple-choice test scores. *European Journal of Psychological Assessment*, *15*(2), 143–150. https://doi.org/10.1027//1015-5759.15.2.143

Roderer, T., & Roebers, C. M. (2010). Explicit and implicit confidence judgments and developmental differences in metamemory: An eye-tracking approach. *Metacognition and Learning*, *5*, 229–250. https://doi.org/10.1007/s11409-010-9059-z

Roebers, C. M. (2006). Developmental progression in children's strategic memory regulation. *Swiss Journal of Psychology*, *65*, 193–200. https://doi.org/10.1024/1421-0185.65.3.193

Schwarz, N. (1994). Judgment in a social context: Biases, shortcomings, and the logic of conversation. *Advances in Experimental Social Psychology*, *26*, 123–162. https://doi.org/10.1016/S0065-2601(08)60153-7

Sherriffs, A. C., & Boomer, D. S. (1954). Who is penalized by the penalty for guessing? *Journal of Educational Psychology*, *45*, 81–90. https://doi.org/10.1037/h0053756

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. https://doi.org/10.1080/13803611.2017.1390479

Swineford, F., & Miller, P. M. (1953). Effects of directions regarding guessing on item statistics of a multiple-choice vocabulary test. *Journal of Educational Psychology*, *44*(2), 129–139. https://doi.org/10.1002/j.2333-8504.1952.tb00885.x

Uncertainty. (1989). In *OED Online*. Retrieved from https://www.oed.com/oed2/00263141;jsessionid=866A4756AFA19B1E760DC5622AFF2269

Waterman, A. H., & Blades, M. (2013). The effect of delay and individual differences on children's tendency to guess. *Developmental Psychology*, *49*(2), 215–226. https://dx.doi.org/10.1037/a0028354