frontiers in Education

ORIGINAL RESEARCH
published: 12 September 2017
doi: 10.3389/feduc.2017.00048

Complement or Contamination: A Study of the Validity of Multiple-Choice Items when Assessing Reasoning Skills in Physics

Anders Jönsson[1], David Rosenlund[2] and Fredrik Alvén[2]*

[1] Department of Science, Kristianstad University, Kristianstad, Sweden, [2] Faculty of Education and Society, Malmö University, Malmö, Sweden

The purpose of this study is to investigate the validity of using multiple-choice (MC) items as a complement to constructed-response (CR) items when making decisions about student performance on reasoning tasks. CR items from a national test in physics have been reformulated into MC items and students' reasoning skills have been analyzed in two substudies. In the first study, 12 students answered the MC items and were asked to explain their answers orally. In the second study, 102 students from five randomly chosen schools answered the same items. Their answers were scored, and the frequency of correct answers was calculated for each of the items. The scores were then compared to a sample of student performance on the original CR items from the national test. Findings suggest that results from MC items might be misleading when making decisions about student performance on reasoning tasks, since students use other skills when answering the items than is intended. Results from MC items may also contribute to an overestimation of students' knowledge in science.

Keywords: argumentation skills, assessment, multiple-choice items, national testing, socio-scientific issues

OPEN ACCESS

Edited by:
Bronwen Cowie, University of Waikato, New Zealand

Reviewed by:
Alison Margaret Gilmore, University of Otago, New Zealand
Robbert Smit, University of Teacher Education St. Gallen, Switzerland

****Correspondence:***
Anders Jönsson
anders.jonsson@hkr.se

Specialty section:
This article was submitted to Assessment, Testing and Applied Measurement, a section of the journal Frontiers in Education

Received: *24 May 2017*
Accepted: *30 August 2017*
Published: *12 September 2017*

Citation:
Jönsson A, Rosenlund D and Alvén F (2017) Complement or Contamination: A Study of the Validity of Multiple-Choice Items when Assessing Reasoning Skills in Physics. Front. Educ. 2:48. doi: 10.3389/feduc.2017.00048

INTRODUCTION

This study investigates the validity of using multiple-choice (MC) items for making decisions about student performance on complex tasks. It has been performed as a reaction to the common practice in Sweden, where most of the national tests include a combination of MC and constructed-response (CR) items.

The national standards in the current Swedish curriculum represent complex skills such as "reasoning skills" (Christenson and Chang Rundgren, 2015) and being able to draw conclusions. For example, by the end of year 6, physics students are expected to be able to discuss questions concerning energy, technology, the environment, and society by posing questions and responding to views "in a way which takes the dialogue and discussions forward." They are also expected to search for information on the natural sciences, use different sources, and reason about the usefulness of the information and sources (Swedish National Curriculum, Lgr 11, p. 125). To interpret, assess, and make decisions about student proficiency in relation to such standards is a complex task. The main purpose of the national tests is therefore to support teachers' decisions when grading individual students according to these standards, assuming that such support improves the fairness and equality of grading.

In order to support teachers in making informed decisions about student proficiency, items included in a test must elicit evidence of the knowledge and skills sought. Such items can be of either MC or CR format. The reasons for including MC items are, for example, objective scoring and internal consistency on the test. Furthermore, since the Swedish teachers assess the tests themselves,[1] the burden of assessment is yet another reason for including such easy-to-score items in the tests. The arguments for including CR items, on the other hand, are often based on validity and alignment with the curriculum (Gustafsson et al., 2014). In the case of complex skills, such as the reasoning skills described above, CR items are sometimes preferred, even if they are more difficult to interpret reliably, because they provide the assessors with direct evidence of students' reasoning. MC items, on the other hand, provide only indirect evidence, since students' actual reasoning is not visible in this format. While MC items may be used for assessing certain kinds of knowledge, the question remains whether they can be used for assessing complex reasoning skills.

To decide whether MC items can contribute to the interpretation of student proficiency in relation to complex reasoning skills, if (and how) students utilize their reasoning skills when solving the MC items needs to be known. The purpose of this study is therefore to investigate the validity of using MC items for making decisions about student performance on complex tasks by investigating student reasoning when solving such items.

In the following section, the relationship between validity and reliability is outlined. Then previous research on students' use of reasoning skills in MC items, as well as the potential interchangeability between MC and CR items, is presented and discussed.

Validity

According to the definition adopted here, whether a test is to be considered valid depends on whether the use and interpretations of the scores are reasonable (Messick, 1996, 1998). This means that validity is not a static property of the test, but depends on different ways of using and interpreting the test results. A basic requirement in the context of national tests, however, is that the tests are aligned to the standards described in the national curriculum. This enables teachers to use the tests to make informed decisions about student performance in relation to these standards.

The question of how to value different contributions to the assessment is sometimes described as a trade-off (Dunbar et al., 1991), where either the validity or the reliability is prioritized. On the one hand, reliability is a prerequisite for validity; otherwise, there would be only static, but no signal. On the other hand, if the tests address something different from what was intended, due to restrictions imposed by improving reliability, the domain to which the results generalize becomes narrower. From this latter perspective, it is not useful to measure how well students are able to identify the correct alternative among a number of distractors, if the test, for example, aims to test students' reasoning skills.

Another way of handling the validity versus reliability trade-off is to attempt to maximize both. One way to accomplish this is to use both MC and CR items on a test. Some items are more closely aligned to the curriculum, but presumably have lower reliability, whereas the opposite is true for other items. Overall, this battery of items might give conditions for both validity and reasonably high reliability. Another way to prioritize both validity and reliability could be to strengthen the reliability for the items more closely aligned to the curriculum, for instance by detailed rubrics, training, and/or applying some kind of moderation procedure. What would *not* be reasonable, however, is to remove all items addressing complex skills or reduce their complexity in order to increase reliability; this would increase reliability at the expense of the validity. This would only work if the purpose and ambitions of how to interpret and use the results from the test are changed simultaneously.

As shown, the relationship between validity and reliability is not a simple dichotomy. Rather, it concerns balancing demands for alignment with the curriculum and levels of certainty in the assessment. The particular question asked here is whether—when attempting to balance validity and reliability—it is meaningful to *complement* CR items with MC items when assessing complex skills. In order to be meaningful in this regard, the MC items would have to provide a valuable contribution to the decisions made from the test results. If the MC items test something other than students' reasoning skills (i.e., the construct is different), they will contaminate and distort rather than complement the possibility to make decisions about student proficiency.

MC versus CR Items

It is not uncommon to assert that MC items can be used to assess students' reasoning skills or their skills in drawing conclusions (Gustafsson et al., 2014). The basis of this argument is that students need to use their reasoning skills in order to tick the appropriate boxes. Even if there is no visible evidence of their reasoning, performance on MC items is considered an indication of students' reasoning skills, similar to CR item responses. This argument is based on some important assumptions: (1) students perform the kind of reasoning that one expects them (in their mind), (2) this reasoning is used to answer the items, and (3) there is an agreement between how students answer an MC and a CR item, addressing the same knowledge or skill. It is only if these assumptions are true that students' answers to MC items are effective indicators of students' reasoning skills when used as a complement to CR test items.

Students' Use of Reasoning Skills in MC Items

In a national evaluation of compulsory schools in Sweden (i.e., not the national tests referred to above), knowledge in science was tested for approximately 3,000 Swedish students in year 9 (the last year of compulsory school in Sweden). A number of the items were MC (see **Figure 1** for examples), and **Table 1** shows

[1] In Sweden, the teachers are responsible for assessing students' performance on the national tests, as well as reporting the results. An "assessment manual" is therefore delivered with the test. This manual specifies the correct answers to MC and other selected-response items, whereas criteria and examples of student responses are provided for CR items. The assessment manual also includes an algorithm for calculating a "test grade" (A–F), either for the entire test or for different parts of the test.

1. Animals consist of a large number of atoms. What happens to the atoms when the animal has died? A. The atoms stop moving. B. The atoms return to the surrounding environment. C. The atoms split up into smaller components which unite and form other atoms. D. The atoms cease to exist when the animal has decomposed.	**2. How far can the piston be pushed?** Johan sucks air into a plastic syringe and closes the opening with a rubber plug as shown by the picture. No air can enter or exit the syringe. The distance from the end of the syringe to the piston is 10 cm (see the picture). Johan holds the rubber plug against a wall (see picture) and tries to push the piston into the syringe. What happens? Tick a box! [Picture of piston, syringe and rubber plug] ☐ The piston cannot be pushed any further. ☐ The piston can be pushed a few millimetres. ☐ The piston can be pushed a few centimetres. ☐ The piston can be pushed several centimetres. ☐ The piston can be pushed all the way to the bottom of the syringe.

FIGURE 1 | Multiple-choice items from a national evaluation in Sweden (translated by the authors).

the frequency of students providing the correct answers. The results reveal that relatively few students were able to provide the correct answer to these items. In particular, only one out of five students provided the correct answer to the right item about the piston. According to Jansson (1994), these results indicate that the common misconception that gases cannot be compressed is very resistant to change; Swedish students have yet to develop a productive scientific model for understanding particles and gases.

However, in the study by Schoultz (2000), the same items were given to another much smaller sample of year 9 students ($n = 20$). Instead of only ticking a box, the students were allowed to answer the questions orally. The students also had access to similar objects (for example, a bicycle pump) to the ones depicted in the items, which the students in the national evaluation did not.

The difference in frequency between the two studies is striking (**Table 1**). Almost all students were able to provide a correct answer to the item with the piston; the same item which very few students managed to solve in the national evaluation. The difference for item 1 is 54 percentage units. It is difficult to draw any conclusions from this comparison, partly because the sample is so different and partly because the whole situation was more

TABLE 1 | Proportions of correct answers to the same items.

	Written performance on multiple-choice item (%)	Oral performance (%)
Item 1	26	80
Item 2	19	90

supportive in the second study. However, even if disregarding the comparison, Schoultz claims that students' difficulties with answering these items had little to do with their science knowledge. Rather, difficulties related to language and interpreting the illustrations led to incorrect answers. Even if the students were able to provide a correct answer orally, in several cases they still chose the wrong alternative for the MC item. Another complication was the contexts of the items. Most were placed in an everyday context, but the students were still expected to provide a scientific answer. According to Schoultz, this change in context was very difficult for the students. This means that the results from the MC items were affected not only by students' language skills but also by their ability to interpret the expectations of the specific testing situation.

In other studies, students' reasoning in relation to MC items has been investigated through "think aloud protocols" (TAPs). This means that students explicated their thoughts while performing the tasks (Hamilton et al., 1997; Reich, 2009, 2015). In the study by Hamilton et al. (1997), high school students completed 16 MC items chosen from a 10th-grade science test. After completing each item, they were asked to summarize and elaborate their reasoning. Given that some of the items were oriented toward mathematical and spatial-mechanical reasoning, they will not be discussed further. However, some items required reasoning skills similar to the ones described in the introduction of this article. Several students in this study used a strategy where they read each alternative, seeking flaws, until they had eliminated all but one option. Whereas some students performed a careful evaluation of each alternative, others chose the alternatives that they thought "made the most sense."

Reich (2009, 2015) used items from the "New York State's Global History and Geography Regents exam," which has proven successful in discriminating between high- and low-performing students. By analyzing students' reasoning during task performance, the results show that students primarily used factual knowledge, reading skills, and "test-wiseness" to solve the tasks, rather than applying the reasoning skills that the test intended to address. Test-wiseness involves skills that can be used to improve test scores, but which are not related to the construct being measured, such as determining the correct option based on how it is formulated.

In yet another study by Dufresne et al. (2002), results from different MC items were compared. They were all intended to test the same well-defined conceptual knowledge (Newton's third law). Although as many as 70% of the students actually chose the correct alternative on one of the items, only one-fourth of those students did so based on a reasoning in line with Newton's third law.

The studies described above are only a small number of selected examples. However, the results from these studies indicate that students may use completely different skills when answering MC items than intended. Representative or not, this means that there is a risk that student performance on MC items might be misleading when attempting to assess complex skills. Since these studies are few, and their respective samples are small due to the qualitative nature of the research, there is a need for further research in this area.

Interchangeability between MC and CR Items

Assuming that a student has a certain knowledge or ability which is relatively constant, he/she should answer all items addressing the same skill at approximately the same level. This is, however, hardly ever the case. Instead, a student with high ability may fail an easy item and *vice versa*. This fact is built into modern models of measurement, such as Item-Response Theory, by estimating the probability for a student with a certain ability to succeed on an item with a specific difficulty (Wilson, 2005; Bond and Fox, 2007). There are many reasons for failing an easy item, despite high ability; some are systematic (such as poor motor skills affecting all written items), whereas others occur randomly (such as occasional misreading). Thus, if one changes the item format from oral to written, one can (on an individual level) expect different results. These depend on both unsystematic factors and systematic factors related to the item format. On a group level, however, it is expected that some of these differences cancel out. For instance, some students are favored by one format or the other.

A number of studies have empirically investigated the differences between MC and CR items. Hogan (1981) conducted a systematic review from more than 60 studies, where the main sample was college students. His conclusion was that MC and CR items may be used to measure the same construct. It is noteworthy, however, that the most common method for investigating whether MC items test the same knowledge or skills as CR items is correlation analysis. Basically, this means that when the correlation approaches 1, one can conclude that the items test the same knowledge or skills.

Serious critique has been raised against the type of studies upon which the abovementioned conclusion is based (see Bennett, 1993). One problem is that the CR items are often reformulated MC items, instead of the other way around. If MC items test different knowledge and skills (such as memory knowledge or test-wiseness) than CR items, the latter constructed from MC items are destined to test the same kind of knowledge and skills as the original items. However, CR items addressing memory knowledge are not of primary concern here, but CR items that address complex knowledge and skills. It would, therefore, make more sense to reformulate CR items into MC items. Another point of critique is that the research is mainly based on correlation analyses, since two items may be positively correlated without necessarily testing the same thing.

Traub (1993) performed a more stringent review of nine studies, which were deemed more methodologically suitable for comparing MC and CR items. Traub's main conclusion from these studies is that there is insufficient evidence to answer the question. However, he does make a distinction between language-related assessment (such as writing ability and reading comprehension)

and assessment in—what he refers to as—more quantitatively oriented subjects (such as mathematics and computer programming). In language-related assessments (three studies), different item formats seemed to test *different* constructs. Different item formats in the assessment of lexical knowledge (two studies) and in quantitatively oriented subjects (three studies) seemed to test *the same* construct. The assessment of reading comprehension (two studies) was more ambiguous, since the results were contradictory.

Generalizing from Traub's review is difficult, partly because it includes so few studies and partly because those studies indicating that different item formats may test the same construct are very limited. For instance, in studies assessing programming, the CR items asked the students to either to make a list of advantages and disadvantages with a certain method or write down a specific procedure for programming. Similarly, in the studies assessing lexical knowledge, students were asked to "fill in the blanks" with single words. This means that the items in Traub's review differ markedly from the examples discussed in the introduction of this article, where the students are supposed to respond to different views and reason about sources. In fact, the correlation between item formats in Traub's review can be explained by the similarities between "recall" and "recognition," which are both examples of memory knowledge (Wainer and Thissen, 1993). The fact that CR items can look so different, in principle they may cover a spectrum from "fill in the blanks" to doctoral dissertations, is a problem when attempting to answer the question about whether MC and CR items test the same construct. However, what is of primary interest is not CR items that address recollection, but rather complex skills such as reasoning.

Results from more recent research in this area also place the conclusion from Hogan and Traub in a somewhat different light. Becker and Johnston (1999) found no relationship between student performance on MC and essay items. Kuechler and Simkin (2004), as well as Bible et al. (2007), found only moderate relationships between MC and CR items. In a meta-analysis by Rodriguez (2003) covering 67 studies, no obvious connection could be found between different item formats. On the contrary, there was great heterogeneity in the material. Rodriguez determined that the construction of the items seemed to affect the connection between item formats. Above all, there was a connection between items with the same kind of item stem.[2] If both MC and CR items have the same kind of stem, they are more likely to test the same construct. The relation becomes stronger if the items also address the same content, but weakens for extended CR items.

In sum, the results from previous research indicate that different item formats are *not* easily exchangeable. Only in some specific cases, when the item stem is similar, can MC and CR items be assumed to test the same construct. Furthermore, this is mainly the case where the knowledge addressed is recollection, and where students' freedom to answer the item is heavily restricted. When students are expected to produce extended answers or perform practical tasks, the correlation drops (Miller, 1998); this occurs either because changing the format also changes the construct or as a result of the lower reliability of the CR items. When trying to assess complex skills, changing from one format to another can therefore be expected to significantly impact the results.

PURPOSE AND RESEARCH QUESTIONS

This study aims to further investigate the relationship between MC and CR items, since current research indicates that: (1) students may *not* necessarily use the complex skills intended when answering MC items, which means that student performance on MC items might be misleading; (2) studies in this area are few and generally have very limited sample sizes; and (3) when investigating the relationship between MC and CR items addressing the same construct, numerous researchers have reformulated MC items into CR items, and not *vice versa*. This may have led to a situation where some aspects of the consequences of replacing CR with MC items have not been investigated. If both item formats measure memory knowledge (and therefore are positively correlated), there is still no evidence of whether MC items are suitable for assessing complex knowledge.

Specifically, this study will analyze students' reasoning skills in physics, where CR items from a national test in science have been reformulated into MC items. The study aims to answer the following questions:

1. When answering MC items constructed to measure students' reasoning skills, which kinds of knowledge or skills are students' reasoning based upon?
2. Is there an agreement between students' answers to MC and CR items designed to address the same reasoning skills?

METHODS

The overall design of this study is to reformulate CR items from a national test, which has been designed to test complex reasoning skills, into MC items and then compare students' answers to the different item formats. By reformulating CR items into MC equivalents, one can assume that the MC items will more likely address the same construct as the CR originals, compared to the other way around. This transformation is described in detail below.

Two samples of students, one for each of the research questions, were asked to answer the MC items. First, a small sample of students answered the MC items during interview conditions, so they were given an opportunity to explain the rationale for their answers. Second, a larger number of students answered only the MC questions (no interviews), so their performance as a group could be compared to a national sample of student performance on the CR versions of the same items.

Transformation from CR to MC Format

The CR items used in this study were taken from the Swedish National Assessment in physics for 12-year-olds. This test typically consists of three parts, all of which are allocated 1 h of

[2]The stem in an MC item is the question or statement that precedes the options.

testing time and which focus on: communication skills (Part A), investigations (Part B), and content knowledge (Part C). Part A is of primary interest, since it includes the assessment of students' reasoning skills. This particular subtest consists of three CR items, each focusing on a particular subskill (Appendix A in Supplementary Material).

First, one task addresses students' skills in using scientific knowledge in discussions about socioscientific issues. For instance, in the 2014 physics test, the one used in this study, the context was the transportation of fruit. The task required the students to make a decision about which mode of transportation to choose for transporting fruit from Italy to Sweden, taking into account environmental concerns.

The second task involves choosing and reasoning about information and sources. In the 2014 physics test, the context was how one's view of the solar system is influenced by technology, religion, and culture. In the task, the students were presented with a number of different sources about the solar system. They were then expected to identify which of these sources were about how (a) technology and (b) religion/culture influence one's view of the solar system. They were also expected to justify their choices (i.e., reason about the information and sources).

The third and final task focuses on using scientific knowledge in order to produce texts, figures, and tables for different audiences and purposes. In the 2014 physics test, the context was the shape of clouds. In the task, the students were presented with a short text about how clouds change during a summer day. The students were then expected to draw a series of pictures, showing how the clouds changed during a summer day and including explanatory captions.

It is important to note that these subtests of communication skills differ from conventional tests in several respects. They are composed entirely of CR items, and they are designed to assess both divergent and convergent thinking. On the one hand, students are free to develop their own line of reasoning. On the other hand, they are also asked to—and rewarded for—using their scientific knowledge to support their reasoning. This means that students can formulate almost any argument they want, as long as they support it (i.e., divergent thinking). However, their support has to be sound. This means that their answers are considered to be higher quality if they use correct and relevant scientific knowledge (i.e., convergent thinking). The more rule-bound and quantitative facet of the physics subject therefore only constitutes one aspect of the assessment. This may also differ from conventional conceptions of tests in physics.

When transforming the CR items into MC equivalents, the first part of each item, which presented the context and any additional information (such as the sources in the second task and the text about clouds in the third), was left unchanged. Therefore, the prompts were identical for both formats. Next, each CR item was divided into two or three MC items in order to reflect the entire breadth of the original item. For instance, in the example above about the solar system, the students were asked to identify relevant sources and justify their choices. When transforming this task into MC items, one item targeted whether the students could identify relevant sources, while a second item addressed whether students could distinguish between appropriate and less appropriate justifications (Appendix B in Supplementary Material). When formulating the stem of the items, care was taken to create similar demands to those in the original item. For example, in the CR version of the cloud task, students were asked to draw a series of pictures, showing how the clouds changed during a summer day. In the equivalent MC items, the students were asked to identify all the kinds of clouds described in the text. In a second item, they were asked to indicate the sequence of changes during the day. Measures were also taken to avoid dependence among items, in the sense that one item cued another.

In order to formulate different alternatives, the assessment manual for the test was used. The assessment manual, which the teachers use when assessing student performance, contains both criteria and authentic examples of student performance for each of the four levels (all items are scored from 0 to 3). Examples of student performance on lower levels were used to create the distractors. Examples from high-scoring students were used to formulate the correct alternative. All alternatives were modified, so that they were similar in length and expression. In total, the three CR items resulted in seven MC equivalents, which were scored as either correct or incorrect.

Table 2 summarizes the characteristics of the MC items. As shown, items 2b and 3b stand out as being more difficult (i.e., have lower *p*-values) and having a greater dispersion. While the discrimination (rank correlation with the sum of all items) of the items can be considered at least fair, the reliability (Cronbach's alpha) is very low. This is expected, partly because there are so few items, but mainly because the original CR items were designed to address different constructs. The correlation between the items is therefore likely to be low. In the table, reliability estimates are grouped for items derived from the same CR item.

Substudy 1

In the first study, schools in the vicinity of the university were contacted to be part of the investigation. Three teachers at different schools answered this request. From each class, the respective teachers selected four students. Teachers were asked to select two boys and two girls, who also represented a combination of high- and low-performing students. Therefore, a total of six boys and six girls with different ability levels participated in the study.

Data consist of students' answers to the MC equivalents. The students took the test individually, and they were asked to explain why they had chosen the particular alternative for each answer (see Hamilton et al., 1997; Shemilt, 2015 for a discussion about this methodology). Students' explanations were recorded with an MP3 player (except for one student who did not want to be

TABLE 2 | Item characteristics for the multiple-choice-equivalent items.

	Item						
	1a	**1b**	**1c**	**2a**	**2b**	**3a**	**3b**
Difficulty	0.55	0.91	0.51	0.68	0.24	0.76	0.33
Discrimination	0.34	0.30	0.54	0.57	0.36	0.41	0.43
Reliability		0.27		0.44		0.19	

recorded, where notes were taken instead). Each session lasted for approximately 20 min. Recordings, which do not include the initial silent reading of information, range from 12:14 to 23:29 min (mean approximately 15 min). Since there were 7 items and 12 students, the entire material consists of 84 choices with accompanying explanations.

Students' oral reasoning was first analyzed by categorizing their explanations as either correct or incorrect and then, independently from whether the explanation was correct or incorrect, on which kind of knowledge or skills they based their reasoning. In the latter case, the ambition was to distinguish between subject knowledge/skills, general knowledge/skills, and test-wiseness, according to Reich (2015). As it turned out, however, the distinction between general knowledge/skills and test-wiseness strategies was not possible to uphold. The reason was that in almost every case the students used general skills as a tool for their test-wiseness strategies. This made it impossible to separate the two categories. Consequently, these two categories were merged and the following categories used:

- Correct answer; no reasoning.
- Correct answer; reasoning based on correct subject knowledge/skills.
- Correct answer; reasoning based on incorrect subject knowledge/skills.
- Correct answer; reasoning based on general knowledge/skills and/or test-wiseness.
- Incorrect answer; no reasoning.
- Incorrect answer; reasoning based on correct subject knowledge/skills.
- Incorrect answer; reasoning based on incorrect subject knowledge/skills.
- Incorrect answer; reasoning based on general knowledge/skills and/or test-wiseness.

The coding was completed by one researcher in relation to explicit and simple criteria. For instance, the students had to refer to scientific concepts in order to be categorized as "based on subject knowledge/skills." The same researcher categorized all student answers on different occasions 2 weeks apart. Any deviations from the initial coding were checked against the criteria.

Substudy 2

In the second study, five schools were randomly chosen from a national database containing all Swedish compulsory schools and contacted to be part of the investigation. In total, these schools had 102 12-year-old students who could take the test. The tests were sent by ordinary (i.e., not electronic) mail to the teachers. They distributed the tests to the students, collected them again, and sent them back to the researchers.

Data for this study consist of students' answers to the MC equivalents, which were scored, and the frequency of correct answers was calculated for each item. The scores were then compared to a sample of student performance on the original CR items from the national test ($n = 7{,}731$; $\alpha = 0.69$). Due to the ordinal nature of these data, a choice was made to use quite crude, but robust statistical tools for analyzing the data.

The national sample comes from teachers voluntarily reporting student performance on the national test through a website, so that they may compare their own students' performance with the performance of all other students reported through the website. This is a service provided by the test developers. The sample in this study corresponds to approximately one-third of all students in the country who took the test.[3] Since the reporting is voluntary and anonymous, no characteristics of this group of students are known, apart from students' test results and gender. This means that it is not known to what extent the two samples are comparable in any other respects. There is, however, no known reason to suspect any bias in the sample.

Ethical Considerations

This study was carried out in accordance with the ethical guidelines for the Humanities and Social Sciences set out by the Swedish Research Council. The study has not been subjected to review by an ethical committee since, according to Swedish legislation regarding research on human subjects (2003:460), research needs approval from an ethical committee only in cases where personal and sensitive information is handled, when physical interventions are made, or when the subjects may be harmed. In line with this, approval from an ethical committee is not required by the university where the research was conducted. All subjects, as well as their legal guardians, have been informed about the purpose of the research, that their participation is voluntary, and that they can interrupt their participation at any time. Written informed consents have been given by all subjects, as well as their legal guardians, in accordance with the Declaration of Helsinki.

RESULTS

Students' Reasoning on MC Items—Substudy 1

Analyzing the answers from the 12 students to the MC equivalents reveals that the students as a group provided correct answers in 69 out of 84 instances (82.1%). Of these 69 correct answers, a total of 31 were based on reasoning using correct subject knowledge. The remaining 38 correct answers were mainly based on a combination of general knowledge/skills and test-wiseness strategies. However, the correct answers may also be based on incorrect subject knowledge/skills. Some students did not provide any reasoning for particular items.

The 15 incorrect answers were almost entirely based on a combination of general knowledge/skills and test-wiseness strategies. However, in a couple of instances, they were also based on subject knowledge. These results are summarized in **Table 3**.

Some interesting observations can be made from **Table 3**. First, approximately the same proportions of reasoning for correct answers were based on subject knowledge/skills and general knowledge/skills plus test-wiseness. These two categories were

[3] In Sweden, schools are randomly assigned one of the science tests. This means that approximately one-third of the students take the test in physics, one-third the test in biology, and one-third the test in chemistry.

TABLE 3 | Overview of students' strategies in answering the multiple-choice-equivalent items.

	Correct answers	Incorrect answers	In total
Subject knowledge	31	2	33
General knowledge and test-wiseness	29	13	42
Incorrect knowledge	6	–	6
No reasoning	3	–	3
In total	69	15	84

TABLE 4 | Students' strategies in answering each multiple-choice-equivalent item.

	Item						
	1a	1b	1c	2a	2b	3a	3b
Correct answers							
Subject knowledge	1	3	5	8	–	6	8
General knowledge and test-wiseness	8	7	4	1	2	5	4
Incorrect knowledge	–	–	2	2	–	–	–
No reasoning	1	2	–	–	–	–	–
Incorrect answers							
Subject knowledge	–	–	–	–	1	–	–
General knowledge and test-wiseness	2	–	1	1	9	1	–

not, however, evenly distributed among the items (**Table 4**). For some items, such as 2a and 3b, students base their reasoning on subject knowledge to a high degree. However, in items 1a and 1b, the students base their reasoning more on general knowledge and test-wiseness. Also, item 2b included many incorrect answers, which were almost entirely based on general knowledge and test-wiseness. Even though all items were very similar, some kind of interaction occurred with either the content or the characteristics of each respective item. Item 2b, for instance, was much more difficult for the students compared to the other items. However, from this small sample of both items and students, it is not possible to investigate the reasons for these interactions.

A second observation from **Table 3** is that for all correct and incorrect answers, the most common base for students' reasoning is general knowledge and test-wiseness. This is due to the fact that it is more common to base incorrect answers on general knowledge and test-wiseness compared to correct answers.

What cannot be seen in the tables, but in the recordings of students' reasoning, is that students basically apply one major strategy for general knowledge/skills and test-wiseness. They compare the different options and reason about the formulations in order to identify the correct alternative. For example, the context for item 1 was the transportation of fruit. The students were asked to make a decision about which mode of transportation to choose for transporting fruit from Italy to Sweden, taking into account environmental concerns. In the following excerpt, a student is explaining his/her answer to item 1c, in which the specific focus was on identifying concerns other than pollution and costs, which were topics covered in items 1a and 1b. As shown, the student reasons by comparing different alternatives, first numbers 1 and 5 and then numbers 3 and 6 (the numbers refer to questions posed by fictional children in the task). This individual finally arrives at number 9 by the process of elimination.

> Yes, I chose number 9, because number 1, which mode of transportation produces the least amount of dangerous emissions, is sort of the same question as here in number 5, what amount of dangerous emissions does each mode of transportation produce per box of oranges. And question 3, which mode of transportation is the most expensive, is also somewhat like number 6, because it is about what it costs to transport a box of oranges with the different modes of transportation. Yes, and then question number 9 is about how much of the oranges it is possible to sell after having transported them with the different modes of transportation. (School 3, Student 1)

This example shows, which is similar across almost all student responses independent of whether they use subject knowledge or not, that students' focus is on both the task and the wording. Similarly, in items 2a and 3b, most students base their reasoning on correct subject knowledge. However, this knowledge is most often used to distinguish between the options, not to reason about the phenomenon regarding the item. This means that although students reason, they do not necessarily provide the reasoning that was intended.

When comparing the options and deciding which one to choose, some notable differences emerged regarding the items. For instance, in relation to items 1a–c, students compare the formulations for the different options and then cross-check with similar wordings or synonyms in the information (remember that information is provided for all items in this test):

> I chose that one because it had the best agreement with the text. (School 3, Student 2)

This means that the students use reading-comprehension skills and word knowledge as general knowledge/skills to find the correct option.

Other items are treated differently. In item 2b, for example, where students are supposed to choose the best justification for a choice of sources, two of the alternatives include the words that students may look for. Nonetheless, the majority of students choose the third (incorrect) option. According to the recordings, this alternative appeals to the students because it is easier to understand, whereas the correct option is more difficult to comprehend:

> /…/Nicklas, I don't even get his justification, but Love, I understand how he's thinking. (School 2, Student 1)[4]

> You understand it well anyway. You know why he chose it… [pause] from the requirements, or what you should call them, up here. (School 1, Student 1)

[4]Nicklas and Love are names of fictional characters who provide the justifications the students are supposed to choose from for the item.

A similar strategy is used in item 3a, where the students choose the most elaborate answer (in this case student drawings), either because it is more elaborate or because it "explains a little bit better":

> Because here things are explained too. On the other hand, there they have only drawn the clouds and not as much is explained, [pause] but this one explains more with the arrows and things like that. (School 3, Student 2)

> Well, it looks like someone has put more effort into this. [pause] It looks like she's spent more time doing this. [pause] Showing how she's thinking and… you know. (School 2, Student 3)

Agreement between Students' Answers to MC and CR Items: Substudy 2

Table 5 compares the score frequencies (actual score relative to the maximum score) on the MC equivalents from 102 students to a sample of students from the national test ($n = 7{,}731$).

Table 5 reveals no obvious pattern, such as the MC equivalents being consistently either more difficult or easier compared to the original CR items. Instead, it differs, so that item 1 seems to be more difficult as a CR item, whereas item 2 is easier and item 3 is of similar difficulty. These observations are confirmed by a median test, showing statistically significant differences between CR and MC items 1 and 2 ($p < 0.05$), but not for item 3. It is also interesting to note the magnitude of the differences. For item 2, it is 12.7 percentage units.

DISCUSSION

The aim of this study was to investigate the relationship between MC and CR items. MC items were therefore constructed from CR originals, which were originally designed to assess students' reasoning skills in physics. The MC items were then answered by 12 students, who also explained the reasons for their answers, and their explanations were recorded. Furthermore, the MC items were sent to five randomly selected schools, so that 102 student answers could be collected. These answers were compared to students' performances (on CR items) on the national test regarding difficulty.

In relation to the first research question, which kinds of knowledge or skills students' reasoning are based on when answering MC items, the findings suggest that students use general knowledge/skills (such as reading comprehension and word knowledge) and test-wiseness strategies (such as comparing the wording of the different alternatives) in the majority of cases. Even when using correct subject knowledge, this is used to distinguish between the different alternatives, rather than being directed toward the scientific context of the task. These findings resonate with previous research in this area, which has shown that students may use different skills than intended when answering MC items. In particular, this study substantiates the research by Hamilton et al. (1997), who showed that students reason primarily about the individual response options by reading the alternatives until they have eliminated all but one. It also substantiates the research by Reich (2009, 2015), who showed that students primarily used factual knowledge, reading skills, and test-wiseness to solve the tasks, rather than applying the reasoning skills the test intended to measure.

The findings from the second study, addressing the question of whether there is an agreement between students' answers to MC and CR items designed to address the same complex skills, also support the interpretation that students may use different skills when answering MC items rather than CR items. The findings indicate that the difficulty for the MC and the CR items differs for two of the three items, despite the effort to make them as similar as possible. Taken together, the results provide indications of MC and CR items not being easily exchangeable. Again, these findings corroborate previous research (Miller, 1998; Rodriguez, 2003).

From the current data, it is not possible to draw any conclusion about the reasons for the observed differences between CR and MC items. However, as a recent study about item difficulty in the Swedish national science test suggests, the difference in difficulty for the CR items can be partially explained by how much of their own knowledge in science the students need to draw upon when answering the items (Jönsson, 2016). In some items, students are provided with all the facts and concepts needed. Their task is to *use* this information by drawing a series of pictures based on a text, as in item 3 of this study. These items are generally easier, in contrast to items where students need to draw upon their own knowledge in order to support their reasoning. Item 1 in this study is one where students need to draw upon their own knowledge, whereas item 2 is intermediate (i.e., some information is provided, but not all). As **Table 5** shows, this gradient in item difficulty exists for the CR items, but not for the MC equivalents, even though the information provided was the same for both item formats.

For instance, the students find item 1 much easier when not having to formulate their own arguments, but instead choosing among different alternatives. A possible reason is that the students do not have to rely on their own knowledge to the same extent. On the other hand, item 2 becomes more difficult in the MC format. As the findings above reveal, the students tend to choose the wrong alternative because they find it easier to read and understand. The explanation for the increased difficulty of item 2 may therefore lie in the fact that students are less familiar with the words and concepts used in some of the alternatives. When answering the same item in their own words, they do not have to rely on word knowledge to the same extent, and item difficulty drops. Similarly, for item 3 the difference is not as significant as compared to item 2. Whereas the CR version of the item requires students to use the text to draw pictures, the MC version relies not only on reviewing others' drawings but also on written text in the alternatives.

TABLE 5 | Mean score for multiple-choice (MC)-equivalent items and the national constructed-response (CR) items.

	MC equivalents (%) ($n = 102$)	CR items (%) ($n = 7{,}731$)
Item 1	65.7	54.2
Item 2	45.5	58.2
Item 3	55.0	61.0

Implications

The inclusion of MC items when assessing reasoning skills in physics may improve reliability estimates, facilitate scoring, and reduce teachers' workload. Nonetheless, the findings from this study suggest that the addition of MC items may *not* necessarily support teachers in making informed decisions about student performance in relation to such skills. On the contrary, the findings indicate that the results from MC items might be misleading; students use other skills than intended when answering the items. Furthermore, since using general knowledge and test-wiseness was one of the main strategies for providing correct answers, MC items are likely to heavily overestimate student knowledge in science (Reich, 2013). Students using other skills than intended could also be true for CR items. However, this is then visible in students' responses to the task. On the other hand, there are indications of the CR items being affected by other construct-irrelevant factors such as drawing and writing skills.

According to these findings, the best way to handle the validity versus reliability trade-off is *not* to combine MC and CR items, but rather to strengthen both the reliability and validity for the CR items more closely aligned to the curriculum. To strengthen reliability, detailed rubrics, training, and/or moderation procedures could be used (Jonsson and Svingby, 2007); to strengthen validity, assessments should not only rely on written responses but also include oral performance.

Limitations and Future Research

Several limitations of this study affect the possibility to generalize the findings. Most important are the items used, since they likely have a significant impact on the results. Specific strengths in this study are that the CR items are designed to address complex skills, and they are thoroughly evaluated with a large number of students, since they are part of a national test. Furthermore, all of the MC items were systematically constructed from the CR originals. All information concerning the task was identical for both MC and CR items, making comparisons more valid. It is not possible to make the MC items perfectly equal to the CR originals, for instance, due to the fact that the CR items were multidimensional (were to be assessed with more than one criterion), and the MC items have to be unidimensional. This means that the MC items could have been designed differently, and other items could possibly produce different results. The number of items (7) used in this study was also small, as an adaptation to the age of the students (12-year olds). Similar investigation, but with other and a greater number of MC items, is therefore a natural recommendation for subsequent research.

Another important limitation to this study is that—like much research in this area—it is small-scale. The first substudy included a convenience sample of 12 students from 3 different schools. Furthermore, the students had to explain the reasons for their answers. This procedure may, on the one hand, have provided more focused data material, compared to, for instance, TAPs. However, it may also have produced the task-oriented answers observed in the recordings as a methodological artifact. Future research investigating students' explanations in relation to CR items or comparisons with TAPs is therefore imperative.

Finally, although a random selection of schools is included, the second substudy is also based on a limited sample of student performance. The sample from the national test is much larger, but is based on teachers' voluntary reporting. Of great importance is the fact that it is assessed by the teachers themselves. Due to the uncertain nature of these data, quite simple statistical tools were used for analyzing the data. More sophisticated methods may have provided more nuance to the findings, for instance, regarding interactions between the students and item characteristics. The final recommendation for future research is therefore to further investigate the statistical relationship of MC and CR items. This should include data that (a) are based on a systematic transformation of CR to MC items, so that the items are designed to address complex skills, but (b) do not depend on teachers' voluntary reporting and potentially unreliable assessment.

ETHICS STATEMENT

This study was carried out in accordance with the ethical guidelines for the Humanities and Social Sciences set out by the Swedish Research Council. According to the national guidelines, as well as Swedish law regarding research on human subjects (2003:460), research needs approval from an ethical committee only in cases where personal and sensitive information is handled, when physical interventions are made, or when the subjects may be harmed. All subjects, as well as their legal guardians, have given written informed consent in accordance with the Declaration of Helsinki.

AUTHOR CONTRIBUTIONS

AJ was the principal investigator, who led the design of the study, literature review, data collection, analyses, interpretation, and writing and revising of the manuscript. DR and FA both contributed to the conceptualization and planning of the study, to analyses, and to the writing and revising of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://journal.frontiersin.org/article/10.3389/feduc.2017.00048/full#supplementary-material.

REFERENCES

Becker, W. E., and Johnston, C. (1999). The relationship between multiple choice and essay response questions in assessing economics understanding. *Econ. Rec.* 75, 348–357. doi:10.1111/j.1475-4932.1999.tb02571.x

Bennett, R. E. (1993). "On the meaning of constructed response," in *Construction Versus Choice in Cognitive Measurement. Issues in Constructed Response, Performance Testing, and Portfolio Assessment*, eds R. E. Bennett and W. C. Ward (Hillsdale, NJ: Lawrence Erlbaum Associates), 1–27.

Bible, L., Simkin, M., and Kuechler, W. (2007). How well do multiple-choice tests evaluate students' understanding of accounting? *Account. Educ.* 17, 55–68. doi:10.1080/09639280802009249

Bond, T. G., and Fox, C. M. (2007). *Applying the Rasch Model. Fundamental Measurement in the Human Sciences*. New York, London: Routledge.

Christenson, N., and Chang Rundgren, S.-N. (2015). A framework for teachers' assessment of socio-scientific argumentation: an example using the GMO issue. *J. Biol. Educ.* 49, 204–212. doi:10.1080/00219266.2014.923486

Dufresne, R. J., William, L. J., and William, G. J. (2002). Making sense of students' answers to multiple-choice questions. *Phys. Teach.* 40, 174–180. doi:10.1119/1.1466554

Dunbar, S. B., Koretz, D. M., and Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Appl. Meas. Educ.* 4, 289–303. doi:10.1207/s15324818ame0404_3

Gustafsson, J.-E., Cliffordson, C., and Erickson, G. (2014). *Likvärdig kunskapsbedömning i och av den svenska skolan – problem och möjligheter [Equity in Assessment in and of the Swedish School – Problems and Possibilities]*. Stockholm: SNS Förlag.

Hamilton, L. S., Nussbaum, E. M., and Snow, R. E. (1997). Interview procedures for validating science assessments. *Appl. Meas. Educ.* 10, 181–200. doi:10.1207/s15324818ame1002_5

Hogan, T. P. (1981). *Relationship between Free-Response and Choice-Type Tests of Achievement: A Review of the Literature*. Washington, DC: National Institute of Education.

Jansson, I. (1994). *Gymnasieelevers kunskaper om materia. En pilotstudie angående de teoretiska linjerna i ljuset av nationella resultat från årskurs 9 [Upper-Secondary Students' Knowledge about Matter. A Pilot Study about the Theoretical Programmes in the Light of National Results from Year 9]*. Studier av naturvetenskapen i skolan [Studies of School Science], Report no. 11. Sweden: Göteborg University.

Jönsson, A. (2016). Student performance on argumentation task in the Swedish National Assessment in Science. *Int. J. Sci. Educ.* 38, 1825–1840. doi:10.1080/09500693.2016.1218567

Jonsson, A., and Svingby, G. (2007). The use of scoring rubrics: reliability, validity and educational consequences. *Educ. Res. Rev.* 2, 130–144. doi:10.1016/j.edurev.2007.05.002

Kuechler, W., and Simkin, M. (2004). How well do multiple choice tests evaluate student understanding in computer programming classes. *J. Inf. Syst. Educ.* 14, 389–400.

Messick, S. (1996). "Validity of performance assessments," in *Technical Issues in Large-Scale Performance Assessment*, ed. G. W. Phillips (Washington, DC: National Center for Education Statistics), 1–18.

Messick, S. (1998). Test validity: a matter of consequence. *Soc. Indic. Res.* 45, 35–44. doi:10.1023/A:1006964925094

Miller, D. M. (1998). *Generalizability of Performance-Based Assessments*. Washington, DC: Council of Chief State School Officers.

Reich, G. A. (2009). Testing historical knowledge: standards, multiple-choice questions and student reasoning. *Theory Res. Soc. Educ.* 37, 325–360. doi:10.1080/00933104.2009.10473401

Reich, G. A. (2013). Imperfect models, imperfect conclusions: an exploratory study of multiple-choice tests and historical knowledge. *J. Soc. Stud. Res.* 37, 3–16. doi:10.1016/j.jssr.2012.12.004

Reich, G. A. (2015). "Measuring up? Multiple-choice questions," in *New Directions in Assessing Historical Thinking*, eds K. Ercikan and P. Seixas (New York, London: Routledge), 221–232.

Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed-response items: a random effects synthesis of correlations. *J. Educ. Meas.* 40, 163–184. doi:10.1111/j.1745-3984.2003.tb01102.x

Schoultz, J. (2000). *Att samtala om/i naturvetenskap: kommunikation, kontext och artefakt [Talking About/in Science: Communication, Context and Artefact]*. Doctoral dissertation, Linköping University, Sweden.

Shemilt, D. (2015). "Commentary. The validity of historical thinking assessments," in *New Directions in Assessing Historical Thinking*, eds K. Ercikan and P. Seixas (New York, London: Routledge), 246–256.

Traub, R. E. (1993). "On the equivalence of the traits assessed by multiple-choice and constructed-response tests," in *Construction Versus Choice in Cognitive Measurement. Issues in Constructed Response, Performance Testing, and Portfolio Assessment*, eds R. E. Bennett and W. C. Ward (Hillsdale, NJ: Lawrence Erlbaum Associates), 29–44.

Wainer, H., and Thissen, D. (1993). Combining multiple-choice and constructed response test scores: toward a Marxist theory of test construction. *Appl. Meas. Educ.* 6, 103–118. doi:10.1207/s15324818ame0602_1

Wilson, M. (2005). *Constructing Measures. An Item Response Modeling Approach*. New York, London: Psychology Press.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.