Threats to Validity when Using Open-ended Items in International Achievement Studies:

Coding Responses to the PISA 2012 Problem-solving Test in Finland

Inga Arffman

Open-ended (OE) items are widely used to gather data on student performance in international achievement studies. However, several factors may threaten validity when using such items. This study examined Finnish coders' opinions about threats to validity when coding responses to OE items in the PISA 2012 problem-solving test. Six discussions during six coder practice sessions (on six OE items) and an interview between five coders were audiorecorded and analyzed by means of content analysis. Three main threats to validity were found: (1) unclear and complex questions; (2) arbitrary and illogical coding rubrics; and (3) unclear and ambiguous responses. Suggestions are given as to how to respond to these threats in order to improve the validity of international achievement studies.

Keywords: open-ended items, coding, validity, international achievement studies

Results of international achievement tests, such as the Programme for International Student Assessment (PISA), are widely used when making educational decisions. However, such decision-making requires evidence for test *validity*, evidence showing to what extent the results lend themselves to such decisions - to what extent the tests measure what they claim to measure - and to what extent the decisions are meaningful. According to Messick (1989, 1995), there are two main threats to validity: construct underrepresentation and construct-irrelevant variance. Construct underrepresentation occurs when a test is too narrow and fails to capture important aspects of what it intends to measure. An example would be a test of mathematical knowledge which only includes arithmetic but not algebra and geometry. Construct-irrelevant variance occurs when a measurement is unduly influenced by factors that are irrelevant to the construct in question. An example is a mathematics test where respondents' performance is unduly affected by their reading and writing ability. Another requirement for decisions made on the basis of a test to be appropriate and meaningful is that it is *reliable*, yielding consistent results over time (test-retest reliability), across different test versions (parallel forms reliability; e.g., different-language versions), and across raters (inter-rater reliability).

Open-ended (OE) and other constructed-response items are often thought to be one way of adding to the construct representativeness and validity of tests, especially when assessing more complex, higher-order processes (e.g., application, evaluation) (see e.g., Lissitz, Hou & Slater, 2012; Livingston, 2009). This is because whereas in multiple-choice (MC) items respondents simply select an answer from a predetermined set of alternatives, perhaps even by guessing, in OE items they need to construct their own answers. OE items therefore form an integral part of current international achievement tests.

However, there are several factors, or sources of error, which may threaten validity when using OE items. One has to do with the *questions* and their wording (Tourangeau, Rips & Rasinski, 2000): questions that contain unfamiliar words and complicated structures cause comprehension problems and construct-irrelevant reading load (Faaß, Kaczmirek & Lenzner, 2008; Johnson, Penny & Gordon, 2009, p. 65), especially for respondents with weak reading skills (e.g., immigrants; e.g., Abedi, 2006); and ambiguously worded questions whose intended, or speaker, meaning (Schwarz, Oyserman & Peytcheva, 2010, p. 180) is not clear and which do not say clearly what kind of answer is required and at what level of detail (Johnson, Penny & Gordon, 2009, p. 84) may elicit unintended responses, masking true knowledge and skills. Another threat to validity related to the questions concerns their complexity: performance on items that require respondents to explain complex phenomena risks being contaminated by writing ability (Lafontaine, 2004, p. 34).

Another major potential threat to validity when using OE item has to do with how responses to these items are assessed and, more specifically, the criteria, or *coding rubrics,* against which they are assessed and coded (for an example, see Appendix 1). To be valid, the rubrics need to measure meaningful aspects of the construct in a fair, reasonable and meaningful way, making clear and meaningful distinctions between performance levels. Vagueness, illogicality and arbitrariness in the scoring rubrics and coding categories thus jeopardize validity. Also, the more performance levels and coding categories there are, the finer the distinctions between them need to be and the harder it is to make these distinctions clear and meaningful (cf. Jones & Vickers, 2011, p. 14-15; Johnson, Penny & Gordon, 2009). Interestingly, it has been claimed that usually in coding rubrics, the main goal has been to ensure clarity, consistency and reliability and that as a result, fairness may have suffered (Huot, 1993; Lumley, 2002; Wu, 2010).

Still a third potential source of error when using OE items concerns respondents and their *responses*. Responses, especially those by respondents with poor language skills (e.g., immigrants and students with disabilities; e.g., Abedi, 2006), may be so erroneously written (e.g., grammatically) that coders cannot understand them and therefore need to code them 0. When this is the case, the true knowledge and skills of the respondent are clouded by his or her writing skills (Weiner, Graham & Naglieri, 2013, p. 117). Or the response may be so vague and ambiguous that coders cannot be sure what the respondent really meant (intended meaning) and whether the response meets the criteria in the coding rubric (Gordon, 1998). This may be because the response is short and leaves a lot unsaid, because it is implicit, or because it contains incorrect and contradictory information (Johnson, Penny & Gordon, 2009, pp. 207-208). In this case, the code given to the response reflects the coder's interpretation of its meaning, which, however, may not be the one intended by the respondent (e.g., Babbie, 2013, p. 263), the code thus providing an invalid picture of what the respondent really can do.

How have international achievement tests managed to cope with the above threats and to ensure the validity of OE items? Not many studies have been made on this. Those that have mainly concern the questions, especially those in English and in older tests (Reading Literacy Study 1991, PISA 2000 and 2003 and TIMSS 1995 and 2003): either the questions have not been clear and unequivocal and have not made it transparent what kind of response has been required (Gutierrez & Ikeda, 2009; Harlow & Jones, 2004; Kwok-Chi Lau, 2009; Ruddock, Clausen-May, Purple & Ager, 2006); or they have contained complex words and structures (Ruddock, Clausen-May, Purple & Ager, 2006); or they have put too much weight on writing skills (Gutierrez &

Ikeda, 2009; Kapinus & Atash, 1994). As a result, students may have underperformed. As for studies concerning coding, Kapinus and Atash (1994) found that coding rubrics in the earliest international tests (e.g., the IEA Reading Literacy Study) were sometimes illogical and unfair, in that they did not measure the intended construct (e.g., reading literacy), but background knowledge, for example. Threats to validity caused by coding rubrics are also implicitly referred to in the paper by Bradshaw (2002; see also the coding guide, e.g., OECD 2000), where she describes steps taken in PISA to ensure inter-rater reliability: she admits that distinctions between codes may be fine and rely on subtle differences in how responses are written; also, partially correct responses may seem to be coded inconsistently and counterintuitively, since some of them receive full credit, others partial credit and others no credit. As for threats to validity caused by student responses, no studies have been made on this. However, that responses may cause validity threat is implied by Bradshaw (2002), who admits that responses are often faulty and hard to interpret, and by many coding guides (e.g., OECD, 2000), which suggest, for example, that responses may sometimes be so poorly written that they "seriously obscure meaning".

More analyses are thus needed to find out how international achievement tests have managed to cope with the above threats and to ensure validity when using OE items. This study examined threats to validity when assessing responses to OE items in the Finnish PISA 2012 computer-based problem-solving test. The purpose was to explore Finnish coders' opinions about threats to the validity of codes they gave to the responses. Answers to these questions help to decide on the accuracy and meaningfulness of decisions made on the basis of the test (especially as far as the Finnish test is concerned) and to develop more valid tests.

METHOD

Test

The test examined was the Finnish PISA 2012 computer-based problem-solving test. PISA is a regular international assessment on reading, mathematics and scientific literacy, administered to 15-year-olds every three years. In 2012, however, also problem-solving was assessed. The assessment contained 16 units and 42 items, all of them new (ie., none were from the previous PISA rounds). Of the 42 items, the vast majority (36) were multiple-choice tasks and/or were coded automatically. Only 5 units contained constructed-response items that needed manual coding, four of them one and one of them two items. Altogether there were thus 6 items that were coded manually. Of these, only one, in a unit called Robot Cleaner, is today open to public (http://cbasq.acer.edu.au/index.php?cmd=toProblemSolving). All other items are still under embargo. Therefore, in this study, whenever possible, examples are taken from Robot Cleaner. When from the other units, they are modified so as to disguise the item. The test was originally prepared in English and French (the source language, or SL, versions), and the Finnish (target language, or TL) version was translated from the English version (for the translation process, see e.g., OECD, 2010).

Coders

In Finland, altogether five coders took part in coding the six items. One of them, who also acted as the coding leader and supervisor, had a PhD in theoretical physics and some experience in

teaching mathematics, chemistry and physics. He was well acquainted with the PISA problem-solving framework and took part in the PISA international coder training session where coders from each participating country familiarized themselves with the coding rubrics and practiced coding with the help of example responses. The coding leader was responsible for selecting and training the other four Finnish coders and checking and analyzing the reliability of the codes. (For more information on the international training, see e.g., OECD 2012a, pp. 39, 45.)

As for the other four coders, the PISA guidelines (OECD, 2012b, 2014, p. 114) say that they should have a good understanding both of secondary-level studies in the relevant subject domain and of secondary-level students and how they express themselves and that they should have a perfect knowledge of the test language. Teachers would thus be expected to be good coders. In addition, however, the coders should be able to commit their time to the coding process for its whole duration, and they should participate in the specific PISA coder training. In Finland, the coders cannot typically be selected from among teachers because of availability issues. Therefore, the four Finnish coders selected were advanced teacher trainees in mathematical subjects, with 3 to 5 years of studies in mathematical subjects and both theoretical and practical training in teaching (e.g., giving lessons and assessing tests). Two had even worked as teacher substitutes. All coders were natives in Finnish and participated in both the coder training and the whole coding process. (In the light of the reliability checks, the coders performed well by international standards; see the section Coder Queries and Reliability Checks, and Table 2.)

Coding Guide

To help coders in the coding process, PISA prepared a coding guide. This contained the general principles guiding the coding (e.g., OECD, 2002) and the coding rubrics for the six items (see Appendix 1). In the general section, coders were reminded that the most decisive factor when assigning codes was validity – whether the student was "able to answer the question". Therefore, for example, spelling and grammar mistakes as well as minor mistakes in numbers were to be ignored. The test was not to be "a test of written expression". Also, it was specified that, for example, responses containing both correct and incorrect information were usually to be coded 0.

The coding rubrics contained the codes used for each item, verbal descriptions of the types of response required for each code and concrete examples of responses for the codes. In half of the items (Items 2, 3 and 5), only two codes were used (1 for Full credit and 0 for No credit); in another half (Items 1, 4 and 6), also partial credit was possible (2 for Full credit, 1 for Partial credit, and 0 for No credit). The rubrics were prepared centrally, at the same time as the items. However, after the international coder training some significant changes were made to them. The rubrics were prepared in English and translated into Finnish.

Coder Practice Sessions

A significant part of the data of the study consisted of six discussions held between the five Finnish coders during six coder practice sessions in May, 2012. Each of the sessions was immediately preceded by a training session (led by the coding leader), during which the coders familiarized themselves with the general principles of coding, the items and their coding rubrics, performed the tasks themselves, and practiced coding by means of international examples of student responses.

During the practice sessions, the coders worked on authentic responses by Finnish students. At first, they tried to code the responses silently and individually, with the help of the coding rubrics and exemplar responses. However, when faced with a response where they were unsure whether the student was able to answer the question, they brought it up for joint discussion. Altogether, they discussed 171 responses (Table 1; in the tables, Item 1 refers to Robot Cleaner). However, of these, only 138 (83.10%) were brought up because of validity issues. The rest were taken up for fun and thus fall outside the scope of this study. When discussing the responses, the coders considered what it was that made them problematic, whether they thought the student was able to answer the question, whether, according to the coding rubric, the responses should be given full credit, partial credit (when appropriate), or no credit, and whether in their opinion these codes provided a fair picture of what the students were able to do.

TABLE 1

The coding leader had planned that about 30 minutes would be devoted to practicing the coding of each item. In reality, however, the time varied greatly between the items, ranging from 3.46 minutes to 41.98 minutes (Table 1). Even though no far-fetched conclusions can be drawn from the number of problematic responses and the time spent practicing, it seems that Item 4, in particular, but also Items 3 and 5 were less problematic to code than the other items. The average practice time per item was 29.17 minutes and the overall time circa three hours. All practice sessions were audiorecorded and transcripted verbatim.

Coder Queries and Reliability Checks

When the coders could not decide themselves how to code a response, they submitted a query to the international consortium, who then provided a solution. In addition, to monitor and check the consistency of the codes given, the consortium calculated reliability indices for the 6 OE items (see OECD, 2014, pp. 258-265). 100 randomly selected responses of each item were coded by all four coders (the other responses were coded by one coder each). To show the level of disagreement between the Finnish coders (and those in all other countries, or locations, too), a coder-item disagreement index was computed. A value of 0 on this index indicates perfect agreement; and the higher the value, the greater the disagreement. On the basis of these indices, national reliability indices were calculated. Finally, the national indices were aggregated across all countries to form an international item reliability index. Items with values under 3 on this index indicate satisfactory consistency, whereas values above 7.5 show high inconsistency.

Table 2 shows the Finnish (national) and the international reliability indices for the six items. In Finland, the indices were consistently much lower (0 to 2.73) than the international indices (1.15 to 4.93), suggesting very high inter-rater reliability (Table 2). In Items 4 and 5 the agreement was complete and in Item 3 it was very high. In contrast, Robot Cleaner caused the most disagreement. All in all, interestingly, these data largely coincide with those in Table 1, suggesting that Items 4, 3 and 5 were indeed not as problematic to code as the other items. The international indices were slightly higher. However, except for Robot Cleaner and Item 6, agreement between the coders was satisfactory.

TABLE 2

Interview

Once the coders had finished coding, a semi-structured joint interview was held with them so as to collect data on the entire coding process. During the interview, the coders were asked about the coder training and practice sessions, the problems they had had while coding, how fair they thought the codes they gave were, and how they would develop coding so as to make it more valid. The interview was conducted by the researcher and lasted about one hour and 20 minutes. The interview was audiorecorded and transcripted verbatim.

Data Analysis

The transcripted practice sessions and interview were analyzed by means of content analysis. The unit of coding was a reference to a threat to validity when assessing OE responses. This could be either one utterance by one coder or several utterances by several coders. A reference could also involve more than one threat (e.g., not infrequently, a threat was related both to a coding rubric and to a response). To classify the references, a coding scheme (together with coding rubrics and code descriptions) was developed. The development was partly deductive, partly inductive: the starting point for the coding scheme was previous research on validity problems with OE questions; however, the scheme was modified in line with data from this study. This led to a coding scheme with three main categories and eight subcategories of validity threats. Coding a subsample of the references showed that the scheme worked relatively well but that the coding rubrics and code descriptions needed some revision. After the revisions all references were coded and classified into the categories. The coding was done independently and

blinded by two coders: the author of this article, and a researcher who has acted as a coder in PISA and other international assessments. Disagreements between our judgments were discussed and resolved by consensus. The agreement rate was 88%.

RESULTS

The coders brought up three main threats to the validity of OE questions: (1) unclear and complex questions; (2) arbitrary and illogical coding rubrics; and (3) unclear and ambiguous student responses. When allowing for both the practice sessions and the interview, by far the most discussed threat was unclear and ambiguous student responses (51.3% of all references; Table 3). The second most common threat was arbitrary and illogical codes (31.6%), and by far the least common threat was unclear and complex questions (17.2%). However, these proportions differed between the practice sessions and the interview (and between the items). This as well as the specific threats to validity caused by the above three factors are discussed in the following.

TABLE 3

Unclear or Complex Questions

During the coder practice sessions, the questions did not cause many problems: altogether, they were only referred to in 11% of the comments (Table 3). Also, questions of two items (Items 3 and 4) were not discussed at all during the sessions, and the question of one item (Item 2) was only mentioned once, suggesting that in these items the questions worked relatively well. In Item 6 in particular, however, the question seemed more problematic. This is supported by the fact

that during the interview when discussing the questions it was especially Item 6 that was mentioned. More generally, too, during the interview problems related to the questions were discussed much more often (30% of all references) than during the practice sessions, suggesting that questions were considered a significant threat to validity. Two interrelated reasons were given for the threats to validity caused by the questions (Table 4): the question was not fully transparent and clear; and the question necessitated complex explaining and writing.

TABLE 4

The question was not fully transparent and clear. In some items, the question was not fully transparent and unequivocal: it did not tell students clearly how specific an answer was required. For example, at times (e.g., in Item 6, a partial-credit item) students were asked to "explain" how they tested a statement. Some students answered as in Example 1 (the examples first give the authentic Finnish response and then its translation into English).

Example 1 Testasin jokaista mahdollista yhdistelmää, joissa annetut ehdot täyttyvät
I tested every possible combination which met the given conditions

The response is in itself correct. However, it is very brief and very general. Therefore, it is impossible to know to what extent the student was "able to answer the question" - had s/he written more. In the coding guide it was specified that to be given credit, the response had to refer to certain variables and combinations between them and that overly general responses would be coded 0. However, students were not provided with this information. They were only told to "explain". The term alone, however, is not enough to make it clear at what level of specificity the explanation had to be.

Also, in the unit Robot Cleaner (Item 1), students were shown an animation on the behavior of a robotic vacuum cleaner (see Appendix 1) and asked to write a rule describing what the cleaner did when it met a yellow block. Many students answered very briefly (Examples 2 and 3).

Example 2 Työntää sen loppuun
Pushes it to the end

Example 3 Se vaihtaa suuntaa
It changes direction

Both answers are, again, correct, but were coded 1, because according to the coding rubric, for full credit (2), two kinds of information had to be mentioned: that the vacuum cleaner pushed the yellow block until it met a wall or a red block; and that the vacuum cleaner then turned 180 degrees. Again, however, students were not provided with this information: they did not know how specific and complete the response had to be. They did not even know that the item was a 2-point task.

Questions necessitating complicated explanations. Another problem with the questions, closely related to the one above, was that to respond to the item, students had to explain complicated phenomena. This, in turn, required that they use complicated language. This problem was more common in partial-credit items (Table 4). For example, when asked to explain how they tested a statement (Item 6), students were to describe relations and combinations between variables, which moreover, were given as fractions. However, some students answered as in Example 4.

Example 4 Testasin jokaista mahdollista yhdistelmää, joissa annetut ehdot täyttyvät
I tested every possible combination which met the given conditions

Responses such as these were deemed to be too general and were coded 0. Again, however, it is impossible to know to what extent the students were "able to answer the question". Why did they respond as they did? If the reason was that they were unable to do the task, 0 was the right code. If they were able to do the task but unable to explain their thinking in writing (or thought that their answer was self-explanatory and adequate as such), code 0 was unfair. In this case, too much weight would have been put on a construct-irrelevant factor, students' ability to express themselves in writing. Interestingly, the coders felt that even they themselves would not have been able to respond to the question in such a way that they would have received full credit, because so much complicated writing was required.

Arbitrary and Illogical Coding Rubrics

As concerns the coding rubrics, the coders' references during the practice sessions and those during the interview differed significantly. During the practice sessions, the rubrics did not seem too significant a source for validity issues, with only 23.3% of the comments referring to them (Table 3). Also, they did not appear to cause problems with each item but only some of them. During the interview, however, the rubrics were by far the most discussed topic: they were mentioned in every other reference. There were two reasons for this, one of them having to do with the timing of the sessions and the interview. Specifically, during the interview the coders also brought up some new problems that had not surfaced during the practice sessions. These were found, not only in the items where problems had been found during the practice sessions already, but also in all other items. Secondly, during the practice sessions, the focus was on student responses, with problems largely looked at from the point of view of the responses;

during the interview, however, the focus was on more general observations about the coding process and its problems. The coding rubrics were what the coders criticized most, (especially those in Robot Cleaner and Item 5). Overall, there were two interrelated reasons for the criticism (Table 5): the codes and distinctions between them were not clear but relied on subtle and arbitrary differences in expression; and the criteria were illogical, counterintuitive and inconsistent. As a result, in the opinion of the coders, the codes often did not provide a fair picture of the students' performance.

TABLE 5

Subtle and arbitrary distinctions between codes. That the codes and boundaries between them were not clear but relied on subtle and arbitrary differences in expression was a very common complaint (the second most common, with 25 references; Tables 4, 5 and 6). For example, in the item in Robot Cleaner, students were required to refer to the cleaner pushing the yellow block until it "met a wall or a red block". Many Finnish students responded by using expressions such as the following (Example 5).

Example 5 Työntää seinää päin - Työntää päin seinää - Työntää kohti seinää
Pushes towards a wall - Pushes against a wall - Pushes towards a wall

The Finnish responses are very similar. The only differences between them are that they use different grammatical words (adverbs "päin" and "kohti"; prepositions in English) and different word order ("seinää päin", "päin seinää"). In contrast, the content words (the verb "push" and the noun "wall"), which typically are the ones that carry meaning in an utterance, are the same. Nevertheless, the responses were coded differently: the first and third were given partial credit

(1) but the second was coded 0. This was because the coding rubric said that for partial credit, the student had to mention either that the cleaner pushed the block or that it turned; but that if s/he specified where the block was pushed to and how much it turned, these had to be "complete and correct". If they were not, the code would be 0. The first and third responses were thus interpreted as being general truths: since there were walls everywhere, the vacuum cleaner always went towards a wall. In contrast, the second response was taken to mean that the vacuum cleaner actually hit the wall, which was not always true. The coders found these distinctions arbitrary and felt that they did not paint a fair picture of what the students were able to do. They did not believe that 15-year-old Finns, when answering the question, gave much thought to what they wrote: what grammatical words they used and in what order. In everyday language, and especially when in a hurry, people typically speak (and write) vaguely, using inaccurate expressions, saying what first comes to mind, not bothering too much about detail.

Illogical, counterintuitive and inconsistent criteria. That the criteria were illogical, counterintuitive and inconsistent was also a common criticism (the third most common, with 16 references; Tables 4, 5 and 6). The problem concerned partial-credit items in particular (Table 5). Thus, at times students did not get any credit, even though their response was partly correct. When asked to write a rule describing the behavior of the robot cleaner (Appendix 1), many students answered as in Example 6.

Example 6 Työntää seinää vasten
Pushes it against a wall

The response was clearly not fully correct, because it did not mention turning. However, it was partly correct in that it did mention pushing. Also, it was partly correct in that in about 50% of

the cases the vacuum cleaner did push the block against a wall. However, the response was coded 0, because where the block was pushed to was not “complete and correct”, as required in the coding rubric. At the same time, however, other partly correct responses (Example 7) were given partial credit (1):

Example 7 Työntää
Pushes

This was because according to the coding rubric, to get 1, the student only had to mention either that the cleaner pushed the block or that it turned. No specification was needed. Short, general and vague responses were thus rewarded, whereas if the answer was longer and more specific, it had to be completely correct to be given credit. In the opinion of the coders, this was counterintuitive and unfair. Usually, at least in Finland, respondents are encouraged to try and write as much as they can. Then they get credit for what they get right. Finnish students may have followed this principle, especially as there was no information in the question as to how the responses would be graded. The coders felt that there was no way of saying which of the students actually knew more.

In other cases, responses were given credit, even though they seemed more or less haphazard and arbitrary. In one item (Item 4), students were given eight numbers, and they were to judge how many of these were needed to complete a task and in which order. To get full credit (2), the student had to give five numbers in the exactly correct order (descending order). However, even responses like the following (Example 8) were given partial credit (1).

Example 8 87654321

The coders felt that students giving answers such as these (all eight numbers in descending order) were not “able to answer the question”. They may even have guessed. Still, as specified by the coding rubric, they were to be given code 1. Compared to this, the partly correct responses coded 0 (e.g., Example 6) were felt to be especially unfair. All in all, the coders felt that the way codes were assigned in the six items given was inconsistent.

Unclear and Ambiguous Responses

Of the three main threats to validity, student responses aroused by far the most concerns during the practice sessions. They were referred to in up to 71.2% of the comments and were the most discussed topic in every session (Table 3). However, during the interview, responses were only mentioned in 20% of the comments and were the least discussed topic. This seeming discrepancy is explained by the different purposes and foci of the practice sessions and interview. During the practice sessions, the coders brought up problematic responses and discussed what it was that made *them* hard to code. The focus was on individual responses. During the interview, however, the purpose was to recapitulate the problems the coders had had while coding. Thus, rather than taking up every problematic response, the coders merely summed up the problems. The main problem with the responses was that they were often so unclear and ambiguous that it was impossible to know what the respondent actually meant. This, in turn, was either because the response was erroneously written; because it used vague and broad words or expressions; because it merely repeated what was said in the question; or because it contained contradictory information (Table 6).

TABLE 6

Erroneously written. A few responses discussed during the practice sessions were so poorly written and ungrammatical that it was impossible to make meaning of them. This was the case, for instance, in Example 9:

Example 9 X on iloinen, jos unta ei ole paljon no kun uni on nyt aik a huoli noin 3 osaa kyllä se on
X is happy, if sleep is not much well cause sleep is now qui te worry about 3 parts yes it is

Because no meaning could be made of the response, it was coded 0. At the same time, however, there was no way of knowing whether the student was "able to answer the question" or not. In cases such as these, a student - usually seemingly an immigrant - was put at a disadvantage because of poor language skills.

Vague and broad expressions. In a much greater number of cases, the response could be understood but the words or expressions used in it were was so vague, broad and ambiguous that the coders were not sure what the student actually meant - what the real, intended meaning of the response was - and whether it met what was required in the scoring guide. This was by far the most common single problem discussed by the coders (34 references in all; Table 6). Consider Examples 10 and 11 (from Robot Cleaner; see also Example 5 above):

Example 10 Työntää palikan reunaan
Pushes the block to an edge

Example 11 Se siirtää sen seinämää vasten
It moves it against a wall (not of a building)

In these examples, it is impossible to know what the students really meant, because the meanings of the Finnish words “reuna” (‘edge’) and “seinämä” (‘wall, not of a building’)” are broad, vague and ambiguous. “Reuna” refers to a spot where *something* ends and something else starts, and “seinämä” to *something* that is like a wall. The students may thus have spoken about a wall (‘an upright side of a building or room’) but they may also have spoken about something else. There is no way of knowing for sure. Therefore, the coders had to make their own interpretation as to what they thought the students meant: If they thought reference was made to a wall, the code was 0; if they thought it was to something else, the code was 1. In the end, responses like Example 10 were coded 1, because “reuna” was interpreted as referring not only to the wall but also to something else. In contrast, responses like Example 11 were coded 0, because they were taken to refer to a wall. However, it is possible that how the coders interpreted these responses is not what the students actually meant. When this is the case, an inaccurate and unfair picture is given of the students’ knowledge and skills.

Reponses whose wording was too close to that of the question. In two of the six items (still confidential; therefore, the example is modified), the problem was that some responses were more or less mere literal repetitions of what was said in the question. For example, in one item students were shown a map with several routes and told that one of them, say, the one from A to B, was impassable but that “all other routes” were OK. Then they were asked whether they would be able to go from spot C to D, and some students answered as follows (Example 12).

Example 12 Kyllä. Koska kaikki muut reitit ovat kunnossa paitsi A:sta B:ään
Yes. Because all other routes are OK except the one from A to F

The response answers the question. Also, it contains both the facts required in the coding rubric

of responses given full credit (1): the answer was “Yes”; and it referred (either explicitly or implicitly) to another link. The problem, however, is that the response only reiterates, or copies, what was said in the question. It does not add anything new to it. Therefore, even though the response does satisfy the requirements for full credit and was coded 1, the coders were unsure whether the student was really able to answer the question and whether the code was a fair representation of what s/he was able to do.

Contradictory information. Sometimes, the meanings of the words used in the response were themselves clear, but the response, in addition to correct information, also contained incorrect information which made it difficult for the coders to decide what the student really meant. This was a relatively common problem especially in those items (still confidential) where, apart from writing a response, students also had to press a button, or where they had to type numbers. Consider Examples 13 to 15 (slightly modified), where students described the functioning of a device.

Example 13	Painaa B-näppäintä niin kauan, kunnes se näyttää 10 Press button B until it shows 10
Example 14	Painaa D-nappia ja kelaa numeroita taaksepäin ja laittaa sitten numero 10 Press button D and wind the numbers backwards and then put number 10
Example 15	Painaa C-painikkeesta niin kauan kunnes siihen tulee 18 Press button C until you see 18

All examples are otherwise correct, but in Examples 13 and 14 buttons B and D (respectively) should be replaced with C, and in Example 15 number 18 should be replaced with 10. However, Examples 14 and 15 were given credit (1), whereas Example 13 was not. The coders thus interpreted that students giving responses such as Examples 14 and 15 actually intended to press C and type 10 (respectively) and were “able to answer the question”; the mistakes were

accidental. In contrast, students giving responses such as Example 13 were interpreted as not being able to answer the question. Actually, however, the differences between the responses are small and there is no way of knowing for sure to what extent each student was "able to answer the question".

SUMMARY AND DISCUSSION

This study examined threats to validity brought up by five Finnish coders when coding responses to OE items in the Finnish PISA 2012 problem-solving test. The coders brought up three main causes of such threats: (1) the question, (2) the coding rubric, and (3) students' responses. The ambiguity of student responses was by far the most discussed threat to validity. It caused a lot of problems in every item (the least in Item 4). Especially common were problems caused by vague, broad and ambiguous words and their interpretation. This was the most common single threat discussed by the coders. Responses containing incorrect information caused problems especially when it was possible that the incorrect information was due to a careful mistake. Grammatically erroneous writing and responses that mainly repeated what was said in the item also caused some problems. Coding rubrics and categories were widely criticized, the main complaint being that distinctions between the categories were not always clear and meaningful but depended on subtle and arbitrary differences in writing. This caused problems in every item. In contrast, illogical and inconsistent coding criteria, which at times sanctioned sincere effort and rewarded guessing, for example, were a problem in partial-credit items in particular. The ambiguity of the questions caused much fewer problems and only in some items (mainly one). Items necessitating complicated explaining and writing were more common in partial-credit

tasks. All in all, the coders felt that the codes they gave often did not provide a fair picture of the students' knowledge and skills. In the following, suggestions are given on how to improve validity when using OE items in international achievement studies.

Improving Validity when Assessing *OE* Responses

Questions. Previous studies show that questions in international achievement tests have not always said clearly and transparently what kind of response and at what level of detail has been required and that many respondents have therefore given too general answers; sometimes, again, so much complicated explaining has been needed to answer the question that those with poor writing skills have been at a disadvantage. (Gutierrez & Ikeda, 2009; Harlow & Jones, 2004; Lau, 2009; Ruddock, Clausen-May, Purple & Ager, 2006). In the light of this study, the above is still true of some items. In addition, this study suggests that more elaborated explanation and writing may be required especially in items where there are more than two coding categories, which, in turn, would make it extremely hard for respondents with weak writing skills to show their true potential in these items, in particular.

More attention thus still needs to be paid to ensuring that questions in international achievement tests are understood as they are intended to be understood (Schwarz, Oyserman & Peytcheva, 2010, p. 180). To this end, questions need to be worded clearly, precisely and unambiguously. For example, it is often not enough just to tell respondents to "explain" something (see also Harlow & Jones, 2004, p. 231); they also need to know how and at what level of detail they are to do this. Besides, this has to be true not only in the SL (e.g., English) but also in all TLs. This is

an extra, huge challenge in international tests and requires, for example, that special attention be paid to making the source instrument translatable so that it can be reproduced in a comparable way in all languages involved (Arffman, 2007, p. 272; Brislin, 1986). Respondents likewise need to be told transparently about the criteria (scoring rubrics) on the basis of which their performance will be assessed. Furthermore, to decrease the amount and impact of writing and to give respondents with weaker writing skills a fair chance to demonstrate their potential, alternative forms of assessment and responding could be used, such as keystroke logging, or capturing all keystroke entries made by respondents and using these data to infer the processes used by respondents while answering. This seems especially important in more complex and elaborated tasks. Also, immigrants, for example, would benefit from the opportunity to respond in their native language (Abedi, 2006). However, accommodations such as these are often laborious and costly to implement. Also, there is probably no form of assessment that would be equally fair to all respondents.

Coding rubrics. Previous studies have also shown that in the first international achievement tests coding rubrics and categories were sometimes illogical and unfair (Kapinus & Atash, 1994). In addition, it has been suggested that an important reason for these illogicalities and unfairness has been that the main focus in the rubrics has been on ensuring consistency (Huot, 1993; Lumley, 2002; Wu, 2010). This study lends some support to this claim, in that even though agreement between the coders about the codes they gave was usually high (Table 2), the codes were often felt to be unfair. This study also suggests that one of the main problems with the rubrics is that distinctions between the categories are not clear but depend on arbitrary and minimal differences in word choice, for example (cf. Bradshaw, 2002). It is important to remember, however, that

this study examined responses by Finnish students, writing in Finnish, (and codes given by Finnish coders) (on the basis of Finnish coding rubrics). It is thus possible that the problem does not concern all languages but only some of them (e.g., Finnish): it is possible that the coding rubrics function as intended in the SL but not in all TLs, because of differences between languages. Another problem with the coding rubrics revealed by this study was that they sometimes run counter to logic and common sense, both that of coders (cf. Bradshaw, 2002) and that of respondents, thus jeopardizing face validity. In addition, this study is in line with previous findings (cf. Jones & Vickers, 2011, p. 14-15; Johnson, Penny & Gordon, 2009) showing that the more coding categories there are, the harder it often is to make distinctions between them logical, fair and meaningful. Again, however, it is possible that this problem does not concern all cultures to a similar extent, because cultures differ in response styles (e.g., Schwarz, Oyserman & Peytcheva, 2010).

More attention thus needs to be paid to ensuring that coding rubrics are, not only easy to code consistently, but also logical and fair, not rewarding guessing, for example, and that distinctions between coding categories are unequivocal and meaningful and do not rely on subtle and arbitrary differences in expression (Moskal, 2003). This, of course, applies, in the first instance, to the SL rubrics, which may need more expert reviews and testing. However, it seems even more urgent to ensure the functioning of TL rubrics in target cultures. The first step to this end is to ensure the translatability and cultural appropriateness (e.g., response styles) of the SL rubrics by involving translators and representatives of various cultures in developing them (Arffman, 2007). More attention also needs to be paid to the translation of the rubrics. At the moment, this receives much less attention than the translation of test items. In addition, not only SL but also

TL rubrics need to be tested sufficiently (e.g., cognitive laboratories). Finally, since increasing the number of coding categories seems to make it harder to keep distinctions between coding categories unequivocal and meaningful and since differences between languages make this even harder, it appears to be wise to use relatively few categories.

Student responses. Even though no systematic studies have previous been made on validity problems caused by responses to OE items in international achievement tests, it has been no secret that students' responses are often erroneously and vaguely written and that it is often hard to know what they really mean (Bradshaw, 2002). However, this study showed how extremely common this problem is: how ambiguous students' responses can be and how often interpretation is required of the coder. The study also suggests that one of the main reasons for the ambiguity is students' often vague, loose and haphazard writing, where vague, broad and ambiguous words are used. Other reasons revealed by this study are grammatically faulty responses, responses that only repeat what is said in the item, and responses possibly containing typing mistakes. Here too, however, it is good to remember that these problems may have been at least partly due to the test having been in Finnish - and its having been delivered by computer.

There is no way of ensuring that written responses are always clear and unambiguous. Students - or any communicators - do not always use clear, specific and unequivocal language, especially when they do not have the time and/or motivation to formulate their message carefully. Also, it is only realistic to expect that responses by immigrants and others with weak language skills will contain mistakes, particularly when they need to write long and complex answers. Likewise, it is to be expected that as assessments become increasingly computer based, the number of

typographical errors will grow. However, it is possible to improve the quality of responses by improving the quality of items - by making sure that the items tell respondents clearly and transparently what kind of answer is required and that, for example, merely repeating what is said in the item is not enough. The vagueness of student writing and the possibility of careful mistakes should also be taken into account in the coding rubrics by ensuring that the rubrics do not rely on minimal differences in word choice or typing. However, the best way to tackle the problem of ambiguous writing would be to use other, less literary forms of responding (e.g., keystroke logging).

Limitations of the Study and Suggestions for Further Research

The most obvious limitations of this study are that only one test in one mode (computer-based), one subject, and one language was analyzed; and that only coders' views were used as data. In order to see to what extent the results of this study are generalizable, comparable studies are needed of other tests, modes, subjects, and languages. For example, more needs to be known about how the use of the computer affects responding and how responding on a computer differs from responding on paper. Also, it would be worthwhile to study whether the significance of writing and word choice increases in reading literacy tests, where connotations, implicit meanings and interpretation typically play an even more important a role. In addition, to see to what extent the threats to validity identified in this study truly affect students' responses, investigations are needed of students' thought processes and response strategies, by means of cognitive laboratories and data logging, for example.

CONCLUSION

This study brought to fore some of the challenges to validity facing international achievement studies. OE items have been used in these studies to fight the threat of construct underrepresentation. At the same time, however, the items have been prone to construct-irrelevant variance and language and communication issues, in particular. Likewise, even though the studies seem to have succeeded satisfactorily in ensuring inter-rater reliability when assessing OE items, guaranteeing the appropriateness, meaningfulness and fairness of the codes given has been much harder. Thus, a major future challenge facing international achievement studies is to increase the validity of OE items and the assessment of responses given to them by decreasing construct-irrelevant variance and the impact of language.

REFERENCES

Abedi, J. (2006). Language issues in item-development. In S. Downing & T. Haladyna (Eds.), *Handbook of test development* (pp. 377–398). Mahwah, NJ: Erlbaum.

Arffman, I. (2007). *The problem of equivalence in translating texts in international reading literacy studies. A text analytic study of three English and Finnish texts used in the PISA 2000 reading test* (Research Reports 21). Jyväskylä: University of Jyväskylä, Institute for Educational Research.

Babbie, E. (2013). *The basics of social research* (6th ed.). Belmont, CA: Wadsworth.

Bradshaw, J. (2008, September). *Ensuring marker reliability in international assessments*. Paper presented at the IAEA Conference, Cambridge, United Kingdom.

Brislin, R. (1986). The wording and translation of research instruments. In W. Lonner & J. Berry (Eds.), *Field methods in cross-cultural research* (pp. 137-164). Beverly Hills: Sage.

Gorden, R. (1998). *Basic interviewing skills*. Long Grove, IL: Waveland.

Gutierrez, R., & Ikeda, H. (2009, November). *Response pattern analysis on the burning candle experiment: TIMSS-based study*. Paper presented at the Third International Conference on Science and Mathematics Education (CoSMEd) 2009 Penang, Malaysia.

Faaß, T., Kaczmirek, L., Lenzner, A. (2008). Psycholinguistic determinants of question difficulty: A web experiment. Proceedings of the 7th International Conference on Social Science Methodology (RC33), Neapel, University of Naples Federico II. Retrieved from http://www.suristat.fr/document/documentArticle/Faass_et_al.pdf

Harlow, A., & Jones, A. (2004). Why students answer TIMSS science test items the way they do. *Research in Science Education, 34*, 221–238.

Huot, B. (1993). The influence of holistic scoring procedures on reading and rating student essays. In M. Williamson & B. Huot (Eds.), *Validating holistic scoring for writing assessments: Theoretical and empirical foundations*, pp. 206-236. Cresskill, NJ: Hampton.

Johnson, R., Penny, J., & Gordon, B. (2009). *Assessing performance. Designing, scoring, and validating performance tasks*. New York: Guilford.

Jones, M., & Vickers, D. (2011). *Considerations for performance scoring when designing and developing next generation assessment*. White paper. Pearson. Retrieved from http://researchnetwork.pearson.com/wp-content/uploads/Performance_Scoring_for_Next_Gen_Assessments.pdf

Kapinus, B., & Atach, N. (1994). Exploring the possibilities of constructed-response items. In M. Binkley, K. Rust / M. Winglee (Eds.), Methodological issues in comparative educational studies: The case of the IEA Reading Literacy Study (pp. 105-133). Washington, D.C: U.S. Department of Education, National Center for Education Statistics.

Lafontaine, D. (2004). From comprehension to literacy: thirty years of reading assessment. In J. Moskowitz & M. Stephens (Eds.), Comparing learning outcomes. International assessments and education policy (pp. 24-45). London: Routledge.

Lau, K. (2009). Critical examination of PISA's assessment of scientific literacy. *International Journal of Science and Mathematics Education, 7*, 1061-1088.

Lissitz, R., Hou, X., & Slater, S. (2012). The contribution of constructed response items to large scale assessment: Measuring and understanding their impact. *Journal of Applied Testing Technology, 13* (3), 1-52.

Livingston, S. (2009). Constructed-response test questions: Why we use them; how we score them. *R & D Connections 11*. http://www.ets.org/Media/Research/pdf/RD_Connections11.pdf

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing, 19* (3), 246–276.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.

Messick, S. (1995). Validity of psychological assessment. Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741–749.

Moskal, B. (2003). Recommendations for developing classroom performance assessments and scoring rubrics. *Practical Assessment, Research & Evaluation, 8* (14). Retrieved from http://PAREonline.net/getvn.asp?v=8&n=14

OECD. (2000). *Coding guide. Prince Edward Island PISA workshop*. Council of Ministers of Education, Canada. Retrieved from http://www.gov.pe.ca/photos/original/edu_PISAcodguid.pdf

OECD. (2010, October). *Translation and adaptation guidelines for PISA 2012*. Paper presented at the National Project Managers' Meeting, Budapest, Hungary. Retrieved from http://www.oecd.org/pisa/pisaproducts/49273486.pdf

OECD. (2012a). *PISA 2009 technical report*. PISA, OECD Publishing. Retrieved from http://www.oecd.org/pisa/pisaproducts/50036771.pdf

OECD. (2012b, February). Procedures for coding paper-based constructed-response items MS12. PISA main study coder training, Salzburg, Austria. Paris: Author.

OECD. (2014). *PISA 2012 technical report*. Paris, OECD publishing. Retrieved from http://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf

Ruddock, G., Clausen-May, T., Purple, C., & Ager, R. (2006). *Validation study of the PISA 2000, PISA 2003 and TIMSS-2003 international studies of pupil attainment* (DfES Research report 772). London: Department for Education and Skills.

Schwarz, N., Oyserman, O., & Peytcheva, E. (2010). Cognition, communication, and culture: Implications for the survey response process. In Harkness, J. et al. (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts* (pp. 177-190). Hoboken, NJ: Wiley.

Tourangeau, R., Rips, L., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge: Cambridge University Press.

Weiner, I., Graham, J., & Nagliewri, J. (2013). *Handbook of Psychology*: Vol. 10. Assessment Psychology (2nd ed.). Hoboken, NJ: Wiley.

Wu, S. M. (2010). Investigating raters' use of analytic descriptors in assessing writing. *Reflections in English Language Teaching*, 9 (2), 69-104.

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing, 27* (2), 147-170.

ROBOT CLEANER

ROBOT CLEANER: Question 3 (Q06)

CP002Q06 – 0 1 2 9

The vacuum cleaner's behaviour follows a set of rules. Based on the animation, write a rule that describes what the vacuum cleaner does when it meets a yellow block.

ROBOT CLEANER SCORING 6

QUESTION INTENT:

Description: Describe the logic governing an unfamiliar system

Nature of Problem Situation: Static

Problem Solving Process: Representing and formulating

Context: Technology, Social

Full Credit

Code 2: Recognises that the vacuum cleaner pushes the yellow block until it meets a wall or a red block AND that it then turns 180 degrees.

- It pushes it until it meets a wall or a red block and then it turns 180 degrees.
- It pushes the block until it meets something else, then it turns around. *[It is not necessary to specify what is met to receive credit. "Turns around" implies a 180 degree turn.]*
- It pushes the block as far as it can, then it turns around (180 degrees). *["As far as it can" implies until it meets something.]*
- It moves it along until it hits something, then it turns completely around and heads back the other way. *["Turns completely around" implies a 180 degree turn.]*
- It turns a half circle after the yellow block can't move any further.
- It pushes until it can't move any further and then it moves off in the opposite direction. *["Opposite direction' implies a 180 degree turn.]*

Partial Credit

Code 1: Recognises EITHER that the vacuum cleaner pushes the yellow block OR that it turns.

- When it meets a yellow block it pushes it.
- Pushes it. *[minimal]*
- Moves it. *[minimal]*
- Turns. *[minimal]*
- Turns 180 degrees.
- Pushes it then turns 180 degrees. *[To receive Code 2, the response must specify that the cleaner pushes the yellow block until it meets a wall or a red block.]*
- The robot pushes the yellow block until it hits something and after that it turns. *[To receive Code 2, the response must specify that the cleaner turns 180 degrees.]*
- It pushes it until it meets a wall or a red block. *[No mention of turning.]*

No Credit

Code 0: Other responses.

- It can't move the yellow blocks.
- It turns a quarter circle. *[If the amount of turn is specified it must be correct.]*
- Pushes it to the closest wall. *[If where the cleaner pushes the block to is specified, it must be complete and correct.]*

Code 9: Missing response (no attempt to answer).

TABLE 1

Descriptive statistics of the coder practice sessions

Session /Item	Student responses			Time spent practicing
	Overall no. of responses brought up	No. of problematic responses	Percentage of problematic responses	Minutes
1	31	27	87.10	38.32
2	42	33	78.57	41.98
3	30	25	83.33	24.96
4	3	3	100.00	3.46
5	24	16	66.67	29.89
6	41	34	82.93	36.38
Average	*28,5*	*23*	*83.10*	*29.17*
Total	*171*	*138*		*c. 180*

TABLE 2
Item reliability indices

Item	*Finland*	*International*
1	2,73	4,93
2	1,33	2,16
3	0,32	1,15
4	0	1,64
5	0	1,30
6	1,83	4,62

TABLE 3

References to the three main causes of validity threats during the practice sessions and interview

	Sessions														Interview		Overall	
	Session 1		Session 2		Session 3		Session 4		Session 5		Session 6		*Total*					
	n	%	n	%	n	%	n	%	n	%	n	%	n	%	n	%	n	%
Question	2	12,5	1	4,5	0	0	0	0	2	12,5	3	23,1	*8*	11	12	30	20	17,2
Coding rubric	5	31,3	6	27,3	0	0	1	50	4	25	1	7,7	*17*	23,3	20	50	37	31,6
Response	9	56,3	15	68,2	8	100	1	50	10	62,5	9	69,2	*52*	71,2	8	20	60	51,3
Total	*16*		*22*		*8*		*2*		*16*		*13*		*73*		*40*		*117*	

TABLE 4

References to threats caused by the questions

	Sessions							Interview	Overall
	Session 1	Session 2	Session 3	Session 4	Sesion 5	Session 6	Total		
	n	n	n	n	n	n	n	n	n
Not transparent, clear	1	1			1	1	4	6	10
Necessitating complex explaining	1				1	2	4	6	10
Total	*2*	*1*			*2*	*3*	*8*	*12*	*20*

TABLE 5

References to threats caused by the coding rubrics

	Sessions							Interview	Overall
	Session 1	Session 2	Session 3	Session 4	Sesion 5	Session 6	Total		
	n	n	n	n	n	n	n	n	n
Subtle and arbitrary distinctions betw een codes	5	5			2		*12*	13	25
Illogical, counterintuitive and inconsistent criteria	4	1		1	2	1	*9*	7	16
Total	*9*	*6*	*0*	*1*	*4*	*1*	*21*	*20*	*41*

TABLE 6

References to threats caused by students' responses

	Sessions							Interview	Overall
	Session 1	Session 2	Session 3	Session 4	Sesion 5	Session 6	Total		
	n	n	n	n	n	n	n	n	n
Erroneously written		1	1		2	1	5		4
Vague and broad expressions	8	10	2		6	3	29	5	34
Too close to the question stem		3				2	5	1	6
Contradictory information	1	1	5	1	2	3	13	2	15
Total	*9*	*15*	*8*	*1*	*10*	*9*	*52*	*8*	*61*