Interrater reliability of constructed response items in standardized tests of reading

Paper for *Nordic Studies in Education*

Introduction

Any test program that relies on human raters to use scales and scoring rubrics in order to judge open-ended item responses needs to be concerned with interrater reliability (Bejar, 2012). For oral presentations, essay writing or extended written responses to reading test items, there are usually no single predefined correct answers. Rather, scoring rubrics must be interpreted by raters and used to determine whether a particular item response displays the expected competence or knowledge. Standardized tests of reading comprehension, such as national tests or the PISA and PIRLS tests, generally include a share of constructed response (CR) items for which this type of rater interpretation of student performance is required. In order to validate the test construction, thus, rating of CR items must be reliable, meaning that raters need to be both consistent and free from different forms of rater effects (Haladyna & Rodriguez, 2013). In short, student scores should depend on the levels of performance rather than on who is doing the scoring. For reasons of ecological validity, in this case the extent to which reading test scores provide plausible and appropriate estimates of the school-based and real-life readings which it proposes to measure, the CR format is often favored by both test constructors and teachers. And although there is mixed evidence with regard to the cognitive demands of different response formats, a share of research points to the fact that CR items may be more apt than for instance multiple-choice (MC) items at measuring various forms of deeper engagement with text (Campbell, 2005; Pearson & Hamm, 2005; Rupp, Ferne & Choi, 2006). According to some analyses of dimensionality in reading tests, CR items also account for a significant and unique share of the variance in reading performance (Kobayashi, 2002; Rauch & Hartig, 2010). However, while CR items may be vital for reasons of ecological validity, the use of it is still restricted in many standardized tests because of problems with rater variation and the MC format is often used instead (Campbell, 2005; Solheim & Skaftun, 2009). Not only may this impede on the test's ability to tap relevant aspects or processes of the knowledge domain, but there is also a lack of research-based estimates of the potentially accessible levels of interrater reliability on open-ended responses. The purpose of this study is therefore to provide more knowledge about both actual and possible levels of interrater reliability in the assessment of reading comprehension and thus better empirical grounds for discussing the development of open-ended test items.

The study of interrater reliability of reading test items is a limited area of research, and the extent of reliability, as well as the exact definition of what might qualify as a "high level" of reliability, will depend on both item construction and on the level of rater training (DeSanti & Sullivan, 1984; Taboada, Tonks, Wigfield & Guthrie, 2013). Therefore, any test program that requires subjective scoring needs to evaluate and validate their own proportion of rater reliability (Bejar, 2012). In the study by DeSanti & Sullivan, seven teacher raters rated test responses to cloze based assessments of reading comprehension. Intra class correlation statistics demonstrated generally high levels of rater reliability across passages, interpretive values and grade levels. In some test designs where extended student responses are scored on polytomous scales, the technical reports often deem interrater reliability to be satisfactory if exact + adjacent agreement extends above 90% (cf. Illinois State Board of Education, 2013). In these cases, large proportions of multiple choice items reassures that the total level of scoring reliability is acceptable. In large-scale testing systems like PIRLS and PISA, several measures are taken to ensure reliable scoring, including compilation of explicit scoring guides for each item and extensive training of raters. In PIRLS, a lower boundary for agreement between raters is set at 85% of exact agreement before scoring of the main data collection can begin (Martin & Mullis, 2012). In PISA [...] (OECD, 2012). Additionally, in programs such as these, scoring reliability is measured not only within countries between the members of the national scoring panel, but also between countries as well as between years.

In the population-based national tests of reading in countries like Norway and Sweden, scoring is generally conducted by class teachers and thus involves a large number of teachers all over the country. Often, teachers score the performances of their own students, or at least students at their own school, which is unusual in a European perspective (EACEA, 2009). Involving a large number of raters means that the level of interrater reliability, as a system potential, is difficult to evaluate before test administration, which means that the quality assurance of test reliability cannot be made in advance, as is the case with PISA and PIRLS. Another consequence is that rater training in order to improve reliability would be an extensive and expensive enterprise, much more complicated than to train the raters of a small panel of experts. Yet, since the national tests measure student proficiency according to curriculum goals, there are good reasons for involving teachers nationwide in the scoring process. Teachers may, for instance, benefit substantially from getting the detailed insight into their students' strengths and weaknesses which the scoring of test responses provide (Brevik, submitted; Wiliam, 2013). It is also likely that teachers by participating in the scoring of national tests and thus being impelled to produce reliable assessments according to national standards will contribute to the reliability of classroom assessments of student performances – something which in the long run is even more important to the equality of assessment at large (Black et al, 2011).

However, a system in which rater training and the improvement of rater accuracy is challenging will need to ensure that open-ended items are constructed in ways that support reliable assessment. This would include, first of all, a careful consideration of the requirements for demonstration of interpretive depth in student responses (Solheim & Skaftun, 2009). Rater variation may depend for instance be related to structural features of items such as the length of the expected response, but also on the cognitive target pursued by a given task. Short-answer questions aimed at assessing the capability of retrieving explicit information in a text may cause fewer problems, since the scoring guide may allow for a high degree of detail in terms of acceptable responses. Items that target interpretive abilities, for instance by asking students to draw conclusions about text meaning on global level or asking them to explain actions or events in a narrative, will most likely exert a greater challenge. For such items, the scoring guideline needs to define the abstracted level of comprehension expected in a large variation of individual test-taker responses. But even the most carefully composed guideline still requires raters to interpret the extent to which a given response matches the intent formulated in the guideline. These interpretations will need to grapple with such classical text theoretical issues as to what extent the construal of text meaning is dependent on the reader's particular position and on the discourse community in which the reading takes place, or to what extent a personally oriented response, based in the reader's previous experiences, may be representative for a more generally defined level or type of reading comprehension.

A related aspect, that may also influence the level of rater variation, is scale length. Commonly, a scale is used to separate responses of different qualities and to provide an opportunity to give partial credit for responses that may not be complete but still not wholly inaccurate. On the one hand, assessing responses on a scale entails that raters make a more detailed use of the information provided in each response. On the other hand, to define multiple levels of comprehension is an interpretive challenge as item difficulty and quality of item responses will appear at several dimensions simultaneously (Cerdan, Vidal-Abarca, Martinez, Gilabert, & Gil, 2009; Mosenthal, 1996; Rouet et al., 2001; OECD, 2009).

In a recent pilot study of interrater reliability in the Swedish national reading test in ninth grade, Author & Author (2016) found that the agreement between raters on open-ended items averaged on .73 (Cohen’s kappa). A commonplace recommendation is that consensus indicators for interrater reliability should be above .80 (Gwet, 2014), yet such benchmarks must obviously be interpreted in the light of the particular purpose and content of the assessment (Kane, 2013; McNamara, 2000). In the Swedish national reading test, a small proportion (25%) of MC items is combined with a larger proportion (75%) of CR items. Rater variation on CR items will thus have a comparatively large impact on the reliability of the test and risk influencing students’ test score considerably. In the study, it was demonstrated that a single student’s test score could vary as much as 12 points – where 66 points was the maximum – depending on who was doing the scoring (Author & Author, 2016). This obviously represents an unacceptably large risk of not being fairly assessed on a test with high stakes for the individual test-taker.

Context of the study

In this study, we examine interrater reliability on open-ended items in the Norwegian national reading test (NNRT) in eighth grade. This test is developed at the Department of Teacher Education and School Research at University of Oslo and administered by the National Directorate of Education and Training (UDIR) in the autumn term of eighth and ninth grade (students aged 13–15 years). The purpose of the NNRT is to assess reading as a basic skill across the curriculum, and thereby provide a means for evaluating the quality of student performance on school and classroom level. As such it provides educators and school administrators with information of learning needs according to competence aims in the national curriculum (KD, 2006, 2013), for instance by detection of students with reading difficulties. The construct definition of reading draws on the three aspects, or reading processes, also used in the PISA test (OECD, 2009): *to retrieve explicit information from the text*; *to interpret and draw conclusions based on information in the text*; and *to reflect on the content of the text* (UDIR, 2015). The composition includes texts (usually ranging between 300–1500 words in length) from various subject fields and 5–7 items to each text. Two item formats are used: standard multiple choice (MC) items, i.e., single correct answer format (Pearson & Hamm, 2005), and constructed response (CR) items. The distribution between the two formats is the opposite of the distribution in the Swedish national reading test with MC items making up approximately 75% of the item sample and CR items 25%. Also in contrast with the Swedish test, where a majority of the CR items are scored polytomously on scales of varying length, the majority of CR items in the NNRT are scored dichotomously as correct or incorrect, yielding 1 point or 0 points. Student responses are typically limited to two lines of text. The scoring guideline to each item provides first of all a generic definition of the correct response and of incorrect responses, and then a list of examples on both correct and incorrect responses.

Up until 2015, the NNRT was a traditional paper-and-pencil test, whereas from 2016 it is a fully digitalized test.

After administration, validity and reliability of the test is evaluated in a technical report published online at UDIR:s official webpage (cf. Author, 2014; NN, 2015). The report includes measures of difficulty over testlets and single items as well as gender differences, internal consistency measures, and detailed results for each item in the test. However, no reports on scoring reliability is provided and to the best of our knowledge there has been no investigation of interrater reliability on the open-ended items in the NNRT since the quality evaluation of the national tests in 2005 (Lie, Hopfenbeck, Ibsen, & Turmo, 2005). At that time, based on a sample from 32 schools, the agreement between the class teacher and an external rater was 90% for items on a dichotomous scale, whereas for items on a three-point scale (0–1–2 points), the agreement was 76% (p. 45).[1] Although the number of open-ended items have been significantly reduced since then, there is still somewhat surprising that interrater reliability is not evaluated continuously, both because it represents a vital aspect of test reliability and because it offers valuable insight into the functioning of individual test items.

Research questions

The present study, therefore, investigates interrater reliability on open-ended items in the NNRT in eighth grade. More specifically, the study pursues the following research questions:

1) What is the extent of agreement between teachers on the one hand and test developers on the other in the scoring of open-ended items in the NNRT?
2) What characterizes items or item responses for which rater variation is comparatively large?

Thereby the purpose of the study is to provide more knowledge about both actual and possible levels of interrater reliability in the assessment of reading comprehension and better empirical grounds for discussing the development of test items and scoring guidelines.

Method

Data sample and participants

The data used for the study included 11 CR items from the NNRT administered in 2015. These items were related to 5 of the 7 texts included in the test and are intended to assess students' ability to retrieve and formulate in their own words meaning which is not explicitly given in the text (UDIR, 2015). The three reading processes are represented according to the distribution displayed in table 1.

Table 1. Distribution of open-ended items over reading processes

	retrieve explicit information	*interpret and draw conclusions*	*reflect on the content*
Item no	2, 8	1, 4, 5	3, 6, 7, 9, 10, 11

Participating teachers (n=20) were recruited from a professional development course on the formative use of national test results. They were asked to rate the open-ended responses of 23 students (253 responses in all), which were distributed to them digitally using the software Questback. All participants volunteered and were informed of the purpose of the study and that their results would be treated anonymously. The participants were all experienced lower secondary teachers (teaching experience ranging between XX and YY years) and all of them had several years of experience from scoring national reading tests.

In the sample of 253 student responses 24 (9.5%) were coded as missing. Since missing data in this case cannot be separated from blanks where students have been unable to provide an answer all missing responses was recoded to 0 (no credit). As a comparison the percentage of missing responses in the whole population of nearly 60 000 students was 9.6%.

A sample of 20 participants is obviously too small to represent the whole population of teachers in Norway who are responsible for the scoring of national reading tests in eighth grade. In terms of investigating interrater reliability of reading assessment among teachers in

Norway, the study should thus rather be treated as a case study and the results will need to be corroborated by future studies using larger and more systematically composed samples of participants. Given the theme of the course from which participants were extracted, it may for instance be reasonable to suspect that the teachers in this study share a particular interest in issues related to reading assessment, a trait that may not necessarily be generalized to the intended population. There is, on the other hand, no apparent reason to assume that the level of interrater agreement in the sample would be very different from the level of agreement in the population. However, in order to investigate the functioning of a particular type of open-ended items in a reading test, and whether the test construction itself generates reliable assessments of the open-ended student responses, the data matrix of ratings (students x items x raters) used in the study is still large enough to produce statistically significant results.

In order to provide comparative data, and to estimate the potentially accessible level of reliability, the study also includes ratings provided by a group of 7 test developers working at the University of Oslo. These participants had all been involved in developing the items used in the study and may be regarded as a sample of expert raters. They were asked to score the same 253 student responses.

In order to answer the first research question, concerned with the extent of interrater agreement between teachers and test developers, we compare rater severity and rater reliability in the two groups and use measures from both classical test theory, such as kappa statistics, and many-facet Rasch modelling. To answer the second research question, concerned with the characteristics of items or item responses for which rater variation is comparatively large, a qualitative item response analysis was conducted.

Classical test theory measures

Cohen's kappa is a consensus estimate concerned with the amount of exact agreement between two raters performing a number of categorical ratings. Thus, the calculations made will represent the distribution of agreement among all the possible pairs of raters (190 pair combinations for the 20 teacher raters and 21 pair combinations for the 7 test developers) including median values. In order to provide a measure for the whole group of raters (teachers and test developers respectively), the analysis also include Fleiss' kappa, which is a reliability measure for the agreement between any number of multiple raters doing categorical ratings on binary or nominal scales (Gwet, 2008; Landis & Koch, 1977). Kappa values are preferred to simply calculating percent agreement because kappa controls for the agreement expected by chance alone (Cohen, 1960). For a binary scale, one would expect that any rater pair would come to 50% agreement just by chance.

In order to interpret indicators of consensus there are several different benchmark values. Landis and Koch (1977) proposed for instance that values between .61–.80 represent substantial agreement, while values above .80 should be regarded as perfect or almost perfect agreement (see also Gwet, 2014). Krippendorff (1980) has argued for a more conservative standard in which values between .67 and .80 should be seen as grounds for tentative conclusions only, while more definite conclusions should require reliability above .80. According to McNamara (2000), "0.7 represents a rock-bottom minimum of acceptable agreement between raters [...] "0.9 is a much more satisfactory level" (p. 58). Irrespective of which of these standards one chooses to confide, it is worth noting that any reliability estimate must be interpreted with regard to the item construction, the scales and the scoring guides used in the particular case and, not the least, with regard to the intended interpretations and uses of test scores (Bejar, 2012; Haladyna & Rodriguez, 2013; Hallgren, 2012; Kane, 2013; Koretz, 2008). Norwegian national test scores are not high stakes for students in terms of immediate implications for grades and future study paths, although the presence of the test themselves is believed to affect the content of instruction, and test results are seen as critical incentives for school development (Skov, 2009; Seland, Vibe & Hovdhaugen, 2013).

Many-facet Rasch measurement

To investigate potential differences between raters in the two groups, data was fitted to a many-facet Rasch measurement (MFRM) model. The basic Rasch model for dichotomous items (Rasch, 1980) rests on the assumption that the probability of a correct answer is a function of test taker proficiency and item difficulty. Thus in its simplest form, the Rasch model can be expressed as:

$$\mathrm{L}n(\mathrm{P}ni/1\text{-}\mathrm{P}ni) = \mathrm{B}n - \mathrm{D}i,$$

where P*ni* is the probability of a correct response by person *n* on item *i*, B*n* is student proficiency for person *n*, and D*i* is item difficulty for item *i* (Bond & Fox, 2015). When B*i* = D*i*, the student has a 50 per cent chance of passing the item.

The MFRM extends the basic model to allow for modelling of the other aspects, or facets, such as criterion difficulty and rater severity. In our case, the model was therefore extended with a rater facet:

$$\mathrm{L}n(\mathrm{P}nij/1\text{-}\mathrm{P}nij) = \mathrm{B}n - \mathrm{D}i - \mathrm{C}j,$$

where the added term C*j* denotes severity for rater *j* (Linacre, 2013). The analysis was made in the FACETS software (Linacre, 2014), which expresses test-taker proficiency and rater severity as measures on the “logit-scale” (logit: log-odds unit). Logits are “non-linear transformations of proportions used to create a linear scale that is more likely to have equal units” (Engelhard, 2013, p. 8). This implicates that the measures are expressed on an interval scale, which in turn enables the analyst to make relevant comparisons of distances between students, items, and raters. By convention, all but one facet is “centered” to have a mean of 0.00 logits.

If the data fits the model, the Rasch-analysis has succeeded to produce invariant estimations of student proficiency, item difficulty and rater severity. If the data does not fit the model, the estimations will not be invariant, effectively limiting the possibility to make meaningful interpretations of the estimates and relationships within any facet, for example the rater facet. Therefore, FACETS reports a so-called fit statistic to each measure. This “quality control” statistics, infit and outfit, indicates the extent to which data fits the model. The expected value for infit (inlier-pattern-sensitive fit statistic) and outfit (outlier-sensitive fit statistic) is 1.0. Values exceeding 1.0 indicate “underfit” (i.e., that a person, item, or rater is unpredictable vis-a-vis the Rasch model). Underfit can result from items measuring a different construct or difficult-to-rate item responses . Values below 1.0 indicate “overfit,” (e.g.,redundant items or range restriction in ratings). As such, overfit is less problematic than underfit. In the presentation below, we will present outfit.

The FACETS output includes a number of interesting reliability statistics (for technical details see Linacre, 2013; Myford & Wolfe, 2003; Schumacker & Smith, 2007). For our purposes, the analysis focuses on the “reliability index,” R, the “separation index,” G, the single rater-rest of raters correlation (SR/ROR), and proportion of exact agreement. R is analog to Cronbach’s alpha for students, and equals traditional test reliability. For raters, R indicates reliability of separation between rater severity measures. Thus, a high R value indicates that the differences between raters would probably be reproduced in another similar rating. R has a ceiling value of 1.0. G has no ceiling value and can be interpreted as the number of distinct groups of test takers, items, and raters in terms of ability, difficulty and severity. Typically, a test designer would like to have large R and G values for students and items, but small for raters. The SR/ROR index indicates to which extent a single rater rates consistently with the other raters (Myford & Wolfe, 2003).

Qualitative item response analysis

In order to identify the characteristics of difficult-to-score item responses, we have used the amount of exact agreement between teacher participants for each of the 253 item responses and analyzed in particular the responses for which rater variation was considerably large. The analysis include considerations of explicitness and preciseness in the response, but also characteristics of the text to which the item relates, item wording, and scoring guidelines. In order to frame our understanding of rater variation in the light of test construction, we also consider the aspect, or the reading process, from which the item is defined, i.e., retrieve, interpret, or reflect. The common traits of the difficult-to-score item responses are discussed by examples from those responses (and items) with the lowest rater reliability.

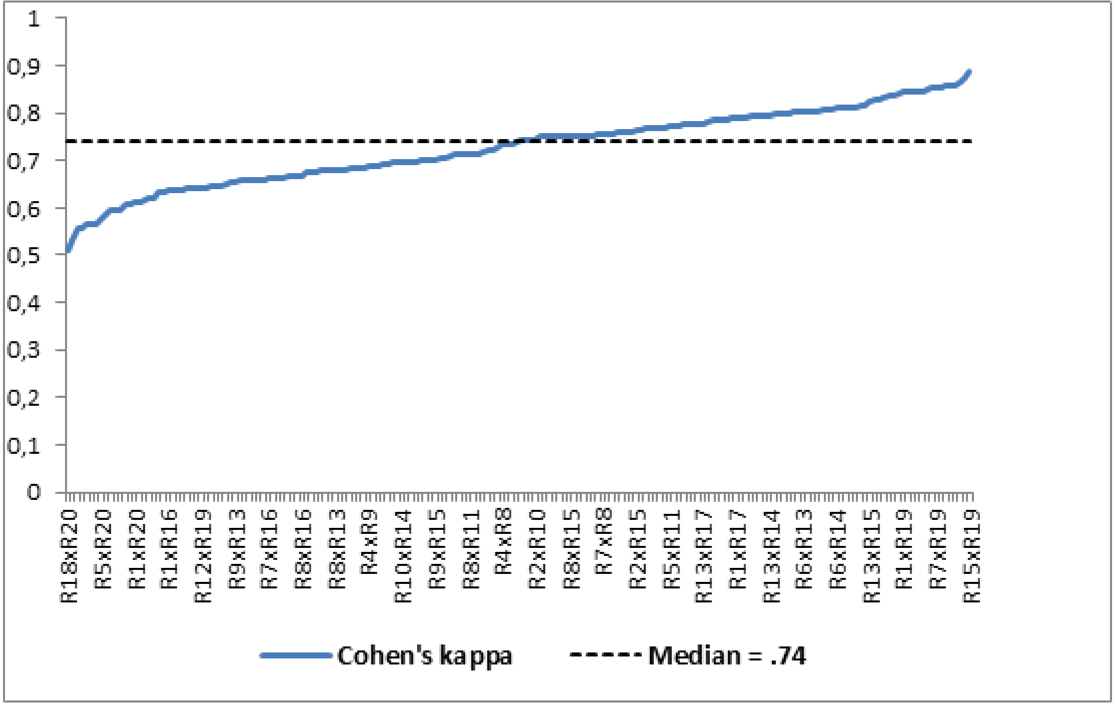
Results

The result section is structured in three sections following the three parts of the analysis. The first section thus reports consensus estimates for the two groups of participants using Cohen's and Fleiss kappa. The second section includes the many-facet Rasch measurement and reports separability and fit statistics as well as rater severity estimates. The third section, finally, presents results from the item response analysis.

Consensus estimates

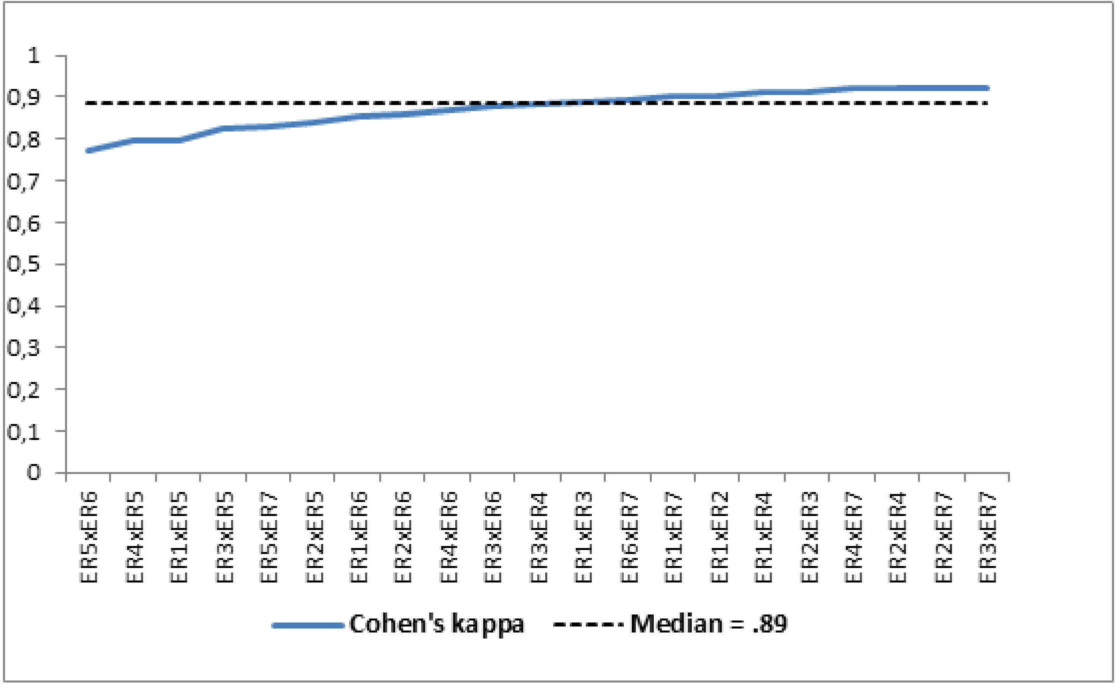
In order to investigate the extent of agreement between raters, Cohen's kappa was calculated for teachers on the one hand and test developers on the other. As noted above, kappa controls for the agreement expected by chance alone and is therefore by necessity lower than if one had made a simple calculation of percent agreement. Fig 1 displays the distribution of agreement between all the 190 pair combinations of teacher raters, showing a spread from a kappa value of .51 to .89 with a median value of .74.

Figure 1. Cohen's kappa for all rater pairs (teachers).



Studying the same statistics for the group of test developers (Fig 2), we can see that the distribution in level of agreement between rater pairs is much more narrow. Cohen's kappa for the pair with the least internal agreement is .77 and for those who agree the most, it is .92. The median value is .89, which according to acknowledged benchmarks (Gwet, 2014; Krippendorff, 1980; Landis & Koch, 1977; McNamara, 2000) should be regarded as quite satisfying.

Figure 2. Cohen's kappa for all rater pairs (test developers).



Since Cohen's kappa can only measure the agreement between two raters at the time, and since a median value is but an approximation of the agreement within the whole group, we also calculated Fleiss' kappa, which, although less recognized or used in the research literature, is the proper measure of reliability between multiple raters (Gwet, 2008; Landis & Koch, 1977). Fleiss' kappa value are reported in Table 2.

Table 2. Fleiss' kappa for interrater agreement between teachers and test developers

alpha = .05	kappa	s.e.	p-value	lower	upper
Teachers	.72	.01	.00	.71	.73
Test developers	.87	.01	.00	.84	.89

//Comment should be added here//

Rasch modelling

The results of the MFRM analysis is presented in a variable map (or Wright map) in Figure 3, and in Table 3. The variable map depicts the logit estimates for each facet. Moving from left to right, the columns represent 1) the logit scale, ranging from -3.0 to 4.0; 2) student scores; 3) item difficulty; and 4) rater severity. Higher scores, more difficulty, and greater severity, respectively, are indicated by higher logit values. Students and items are denoted by number, while raters are denoted by numbers and “T” for “teachers” and “TD” for “test developers”. As expected, there is a noticeable spread in measures for both student ability (distributed over the whole scale) and item difficulty (between -0.89 and 2.42). However, there is also a spread in rater severity (ranging between -0.33 to 0.73). When separating teachers from test developers, as in figure 4 and 5, we can see that the spread among teachers, with values ranging from -0.33 to 0.73 logits, is substantially larger than for test developers (ranging from -0.12 to -0.28 logits).

Figure 3. Variable map including student ability, item difficulty, and rater severity.

```
+------------------------------------------------------------------------+
|Measr|+Students        |-Item   |-Rater                                  |
|-----+----------------+--------+----------------------------------------|
|   4 +                +        +                                        |
|     |                |        |                                        |
|     | 13             |        |                                        |
|     |                |        |                                        |
|     |                |        |                                        |
|     |                |        |                                        |
|     |                |        |                                        |
|   3 +                +        +                                        |
|     |                |        |                                        |
|     |                |        |                                        |
|     | 21             |        |                                        |
|     |                | 1      |                                        |
|     | 19  4          |        |                                        |
|     |                |        |                                        |
|   2 +                +        +                                        |
|     | 6              |        |                                        |
|     |                |        |                                        |
|     | 20             |        |                                        |
|     | 9              |        |                                        |
|     | 15             | 7      |                                        |
|     | 5              |        |                                        |
|   1 + 12             +        +                                        |
|     |                |        |                                        |
|     | 1              |        | T20                                    |
|     |                |        | T12  T18  T9                           |
|     | 14  8          |        |                                        |
|     | 11             |        | T16                                    |
|     | 16  17  3   7  | 5   9  | T14  T17  T3   T6   T7                 |
*   0 * 10             * 8      * T1   T8                                *
|     |                |        | T10  T15  T2   T4   TD1  TD2  TD3  TD6 |
|     | 2              |        | T11  T13  T19  T5   TD4  TD5  TD7      |
|     |                | 10     |                                        |
|     | 22             | 11  2  |                                        |
|     |                | 4      |                                        |
|     | 18             | 3   6  |                                        |
|  -1 +                +        +                                        |
|     |                |        |                                        |
|     |                |        |                                        |
|     |                |        |                                        |
|     |                |        |                                        |
|     |                |        |                                        |
|     |                |        |                                        |
|  -2 +                +        +                                        |
|     |                |        |                                        |
|     |                |        |                                        |
|     |                |        |                                        |
|     |                |        |                                        |
|     |                |        |                                        |
|     |                |        |                                        |
|  -3 + 23             +        +                                        |
|-----+----------------+--------+----------------------------------------|
|Measr|+Students        |-Item   |-Rater                                  |
+------------------------------------------------------------------------+
```

Figure 4. Rater severity among the teachers.

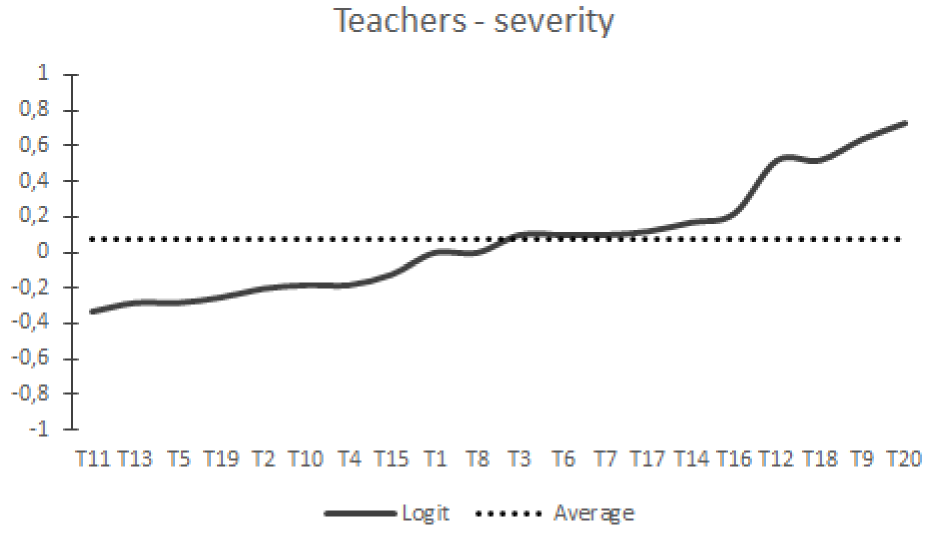


Figure 5. Rater severity among the test developers.

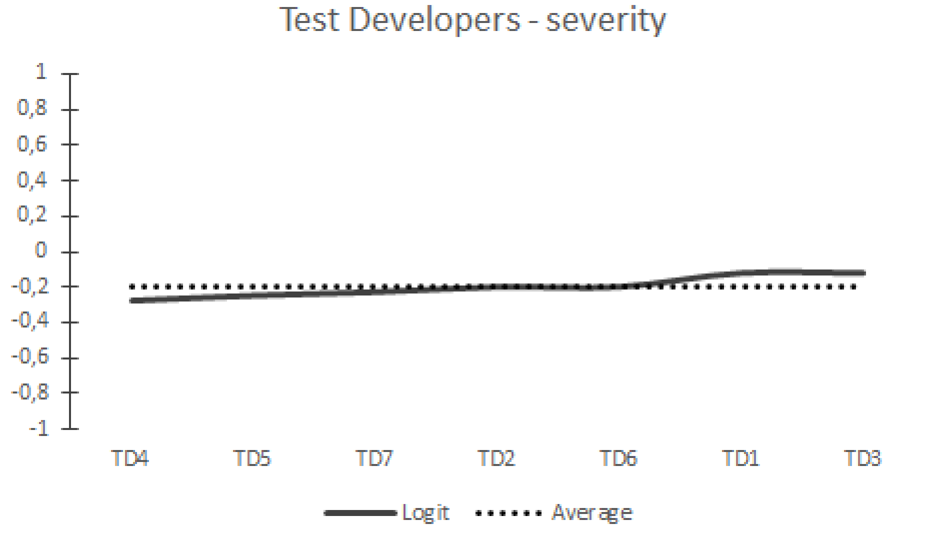


Table 3 provides summary statistics for each facet. The significant chi-square as well as high R- and G-values for students and items, respectively, indicate high precision measures and that the difference between students and items were not by chance. For test developers the chi-square was non-significant, and R- and G-values < 0.01, which indicate non-measureable differences between the raters. In other words, they functioned interchangeably. For teacher raters, there was a significant chi-square, with an R-value of .76 and a G-value of 2.70, indicating at least two distinct groups of severity. The average correlation between single rater and rest of raters was .31 for teacher raters, and .34 for expert raters, a finding that corroborates the separation results.

Table 3.

	Students	Items	Raters	
			Experts	Teacher

Logit spread	6.72	3.31	0.16	1.06
Chi-square	1072 (22)**	939.8 (10)**	0.8 (6)	81.0 (19)**
R-value	.99	.99	.00	.76
G-value	8.37	10.06	0.00	2.70
% agree	-	-	94	86.9
Outfit > 1.5 (%)	4.3	9.1	0	10

** p < .00

According to the measures of outfit, one of the students and one of the items (item 4, outfit 2.20) demonstrated significant high outfit (student 19, outfit 1.74 and item 4, outfit 2.20). We also see that two of the teacher raters demonstrated significant outfit (T9, outfit 1.50; T12, outfit 1.61). None of the test developers demonstrated out of range outfit values.

Summing up, the results show that the rater group as a whole produced ratings that fitted the MFRM model well. This can be attributable to effective scoring guidelines and a good match between item difficulty and student proficiency. Separating experts from teachers, however, revealed non-trivial differences in severity. While the test developers could be used interchangeably, the same is not true for the teachers.

Characteristics of low reliability item responses

In order to locate potential sources of rater variation, and to illustrate in more detail the particular challenges faced by teachers in the scoring of open-ended responses, we will now describe the qualitative characteristics of responses and items on which agreement between raters was comparatively low. When identifying these items, we concentrate on ratings performed by the teacher participants.

With 20 participants scoring each response on a dichotomous scale, there will be 11 possible degrees of agreement (see table X). As shown in the table, the teachers were in complete agreement on more than half of the responses, and nearly four of five responses (79.4%) obtained 85% agreement or more.

Table 4.

Agreement	**Number of responses**	**Percentage of responses**
20–0 (100%)	133	52,6
19-1 (95%)	37	14,6
18-2 (90%)	17	6,7
17-3 (85%)	14	5,5
16-4 (80%)	13	5,1
15-5 (75%)	11	4,3
14-6 (70%)	9	3,6
13-7 (65%)	8	3,2

12-8 (60%)	8	3,2
11-9 (55%)	2	0,8
10-10 (50%)	1	0,4

Amount of agreement may also be inspected for individual items. Figure X displays the responses to each item distribution into three levels of agreement: total agreement (100 %); substantial agreement (95–80%); and low agreement (< 80 %). Items are numbered 1–11 according to their order in the booklet.

Items 1, 2, and 8 include responses which obtained total or substantial agreement only. In item 1 the students were asked to name three linguists, and although they had to make an inference to find the information in a footnote, the linguists' names were explicitly stated. Item 2 asked for a certain part of the country, which was also explicitly stated in a diagram. Item 8 also asked for specific information, here a cartoon genre, which the students either mentioned or not. The reason we find disagreement for 6 responses here is that one coder systematically misinterpreted the scoring guide. Hence, when the information needed to solve the task is explicit in the text and the scoring guide is unidimensional, reliable scoring of an open-ended response seems to be easy.

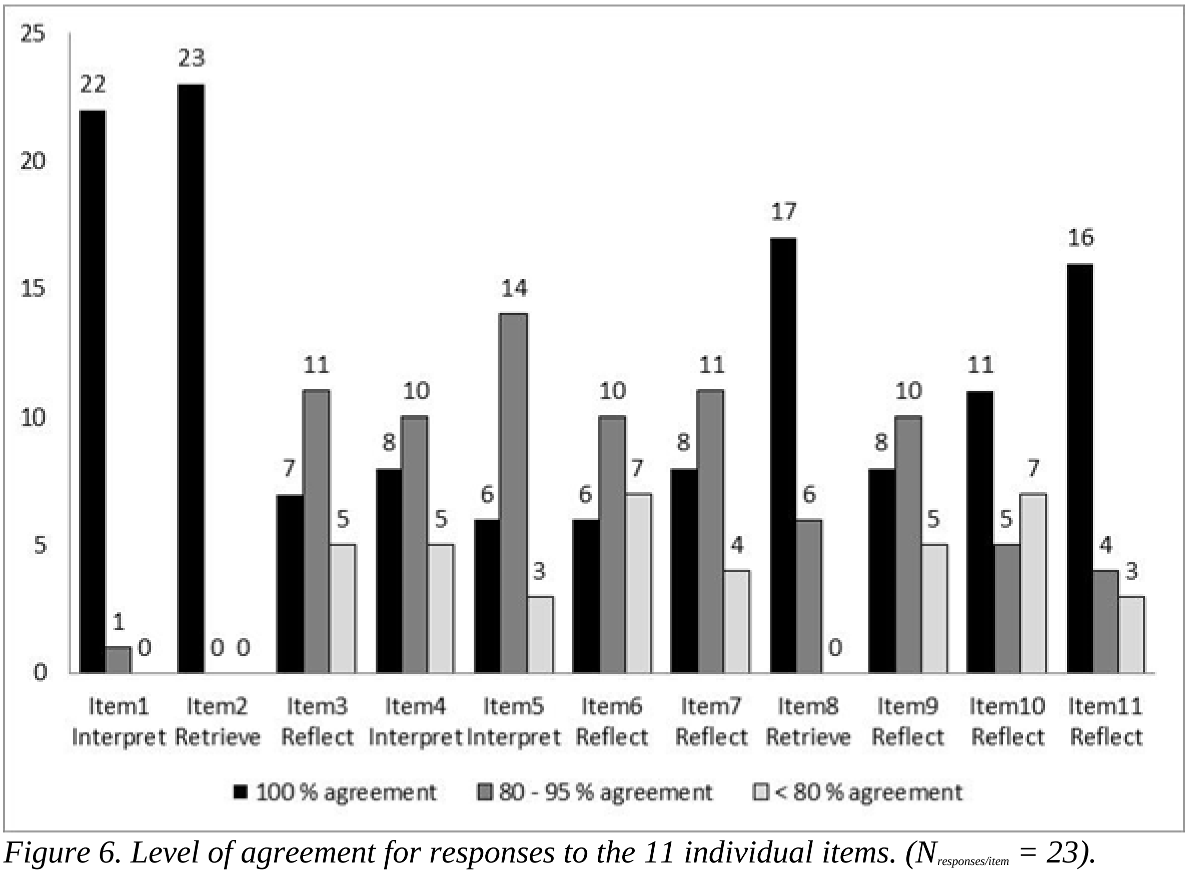


Figure 6. Level of agreement for responses to the 11 individual items. ($N_{responses/item}$ = 23).

To items 3, 4, 6, 9, and 10 there were at least 5 responses that resulted in low agreement (< 80%). Different from items 1, 2, and 8, these items ask for explanations of the content and form of the text. In order to solve the task, the test-taker will have to provide, for instance, reasons that explains a character's emotional state, or reasons for why certain information is provided (or not provided) in the text. Hence, the rater will have to judge not only the accuracy of the interpretation itself, but also whether the written response contains enough adequate details to demonstrate comprehension according to the scoring guide.

As evident from table 4, a number of responses obtain particularly large disagreement. We will now look more closely into the ones on which agreement is 60% or less. These 11 responses are displayed in table 5 along with item wording and scoring guide.

Table 5. Responses with agreement of 60% or less.

Item and aspect	**Scoring guide**	**Student response**	**Comment**	Reliability % acc.	
				T[a]	TD[a]
4. *What surprises the main character at the beginning of the text?* (interpret)	Refers to the fact that the father breaks his usual routine, OR that he does something that the main character does not expect.	What surprised him was that his father wanted to go skiing when he didn't know how to do it.	Not explicitly stated in the text whether the boy knew that his father was a bad skier or not (see also below).	50	86 (6-1)
		What surprised him was that the doorbell rang	Insufficient, lacks an explanation of why the ringing of the doorbell surprised him.	60 (12-8)	86 (6-1)
6. *In the last sentence it says that the main character was relieved that the skiing trip was over. What may be the reason?* (reflect)	Refers to the main character who finds the skiing trip embarrassing, humiliating for the father, OR who was tired of pretending not being able to ski.	He did not think it was fun to go skiing with his father.	Two possible interpretations of "he did not think it was fun". 1 He didn't enjoy it. 2. He was uncomfortable with it	55	57 (4-3)
7. *The book got a lot of attention when it first came out in 1954, Why has it become a collector's item?* (reflect)	Refers to the book being a part of the cartoon history, OR to its impact on the development of cartoons.	He thought that was the reason for youth crime because it was so violent, and it has become a collector's item because cartoons stopped having so much violence in in them	The student presents two explanations, one is correct, the other is irrelevant. All TD:s gave credit	55	100
		Because he wrote useful things about cartoons and how they make them	The response touches vaguely upon a plausible explanation. TD:s did not give credit.	60	100
		Because he made horror magazines disappear from the market	The student indirectly gives a correct explanation. TD:s gave credit	60	100
9. *The text is about cartoons, why are different mass media included in the diagrams?* (reflect)	Refers to the fact that other media are included to compare them with the use of cartoons.	To give us more information. And that you can read cartoons on the internet and in newspapers as well.	The response does not compare cartoons with other media (see also below)	60	71 (5-2)
		To get a better overview of what has gone up and down	Minimal response that suggests a comparative aspect (see also below).	60	100
10. *Most of the graphs go from 1990 to 2012, but two of them are shorter, What may be the reason?* (reflect)	Refers to the two graphs and that they represent new media, OR that there is no data or information about these media	That the thing disappeared or was invented then.	The response provides two explanation of which one is vaguely related to a correct response.	60	86 (6-1)

	earlier.	Because it came late.	Short, minimal, implicit answer, experts gave credit and benefit of the doubt.	60	100
11. *What may be a reason why some people get feel threatened and become aggressive when they visit Hundertwasserhaus?* (reflect)	Refers to a negative view of the house, i.e. that it is ugly, unpractical unusual OR that the inhabitants have too much freedom to do what they want.	Because he has trees inside and so on.	Short but sufficient and a good example. Experts gave credit.	60	100

[a] T = Teachers, TD = Test developers

As shown in table 5, 9 of the 11 responses that caused substantial rater variation are connected to reflect-items and two are connected to an interpret-item. Common to these responses is that they are vaguely worded or insufficient in terms of relevant details from the input text. Thereby, they require interpretation on behalf of the rater, through which variation in severity will impact the scoring. While some raters will give benefit of the doubt, as they are instructed in the scoring guide, others will use vagueness as an indication of limited comprehension. It is also interesting to note that for types of responses not found exemplified in the scoring guide, the teachers give credit to a lesser extent than the test developers. This indicates that the number of examples provided may influence the reliability of scoring..

In the following, we analyze in more detail the characteristics of three of the responses above. One of them was provided to an item related to a narrative text, while the two other were given to an item related to a descriptive text.

Item 4, student 19

This response caused total disagreement between the teacher participants: ten gave credit while the other ten did not. Among the test developers, however, six of seven gave credit to this response. The item relates to an excerpt from a novel, portraying a difficult father-son-relationship. It begins one morning when the doorbell rings. The son opens to see his father dressed in new ski equipment, announcing that the two of them are going skiing together. Eventually we learn that the son is a good skier, but the father is not. The son tries to save his father from embarrassment by hiding his own skills and pretending not to notice his father's weaknesses.

Item 4: What surprises the main character at the beginning of the text?
(Aspect: Interpret and draw conclusions)

Scoring guide:
Full credit (1): Responses refer EITHER to the fact that the father broke his routine, OR that the father did something unexpected, for example:

- The father does not just go into his house, he rings the doorbell.
- The father invites the boy on a skiing trip.
- The father has bought a new skiing equipment.

No credit (0): Responses refer to something that happens later in the text OR vague, incomplete and irrelevant responses, for example:

- The father wanted to go skiing near the military camp. (happens later)
- The father didn't know how to put on his skis. (happens later)
- The doorbell rang. (not surprising in itself, vague, incomplete)

The student response:
"What surprised him was that his father wanted to go skiing when he didn't know how to do it."

10 teachers gave this response full credit and 10 teachers did not, whereas only one of the test developers gave 0. In the text, it is not clearly stated whether the son already knew that his father was a bad skier, or whether he experienced this after they had started skiing later that day. This may be one cause for the disagreement. If he already knew, the surprise would be natural, and the response should be given credit, but if not, there would be no surprise at that point and therefore no credit for the response. 15 of the 23 student responses to this item caused some level of disagreement between the raters, which indicates that the item itself may be problematic. Items 5 and 6, connected to the same text, also resulted in a fair deal of disagreement between raters (see figure X).

Item 9, student 3 and 13
The second example is from item 9 where two responses ended up with a reliability on 60%. The item is connected to a text about the history of cartoons, including two line charts which present boys' and girls' use of seven different mass media from 1990 to 2012.

Item 9: The text is about cartoons, why are different mass media included in the diagrams?
(Aspect: Reflect on the content)

Scoring guide:
Full credit (1): Responses refer to the fact that the other media are included to compare the use of these with the use of cartoons, for example:

- To show what teenagers do today. (Implicit comparative aspect)
- To compare today's media.
- To show that it has become more internet than reading.

No credit (0): Responses do not compare cartoons with other media, but only refer to the development over time, OR vague and irrelevant responses.

- Because it would have been impossible to make the diagram. ("correct" but irrelevant)
- To show that there were more children who read cartoons earlier. (no comparison with other media)
- To show how cartoons have decreased and increased. (no comparison with other media)

The two student responses:
Student 3: "To give us more information and that you can read cartoons on the internet and the in newspaper as well."
(8 teachers scored 1 and 12 teachers scored 0; 2 test developers scored 1 and 5 scored 0).

Student 13: "To get a better overview of what has gone up and down."
(12 teachers scored 1 and 8 teachers scored 0, all 7 test developers scored 1).

Unlike the case with the narrative, there is little ambiguity in this text, and facts are clearly presented in the diagrams. The scoring guide emphasizes the comparative aspect. The item, however, opens for several different ways of reasoning.

The first response does not include a comparative aspect, although it is an undisputable fact that one can read cartoons on the internet. Therefore, while it isn't wrong, it fails to explain the content of the diagram. The explicit reference to the internet may, however, have been interpreted as a comparative aspect. Even 2 of the test developers gave credit to this response.

The second response can be interpreted as suggesting a comparative aspect by stating that something has gone up, and something else gone down. The fact that all seven test developers gave credit to this response indicates that this implication has been taken into account in the rating. Presumably, they also adhere to the general introductory advice to give benefit of the doubt in cases of uncertainty. On the other hand, the response is clearly vague. While the examples of acceptable responses in the scoring guide all refer either to teenager habits or media use, this response has no reference to the content of interest. Teachers who score 0, may therefore have interpreted it as pointing to no comparison between cartoons and other media.

Also for item 9, 15 responses resulted in disagreement between raters.

Discussion

The overarching purpose of the study has been to provide more knowledge about rater variation in the assessment of reading comprehension and thereby better empirical grounds for discussing the development of test items and scoring guidelines. To fulfil this purpose, the NNRT was used as a case for investigating 1) the extent of agreement between both teachers and text developers in in the scoring of open-ended items, and 2) the characteristics of items or item responses for which rater variation was comparatively large. The results of the study indicate that while the test developers produced reliable ratings according to both classical test theory measures and MFRM, the teacher participants varied considerably in terms of both consistency and consensus. Based on closer inspection, we found that substantial variation was obtained specifically for responses to interpret and reflect items and for responses that were vaguely worded or insufficient in terms of relevant details from the input text.

These results raise some crucial concerns for researchers as well as for test developers and school administrators. First of all, it needs to be considered whether the estimated levels of reliability should be regarded as a problem that requires action. Second, if actions are required, what sort of system qualification would then be both justifiable, in terms of cost effectiveness, and practically available?

As noted above, there are few studies available that report levels of agreement between teachers or expert raters on open-ended items. Interestingly, the consensus estimates for scoring reliability in both the Swedish and Norwegian national reading test end up close to .73 (c.f. Author & Author, 2016), although both item construction and the structure of scoring guidelines differ substantially between the two tests. However, while this level of rater variation may have a large impact on students' test results in the Swedish case, since open-ended items are in majority and since the rating scales are polytomous, the range of possible variation of student results in the NNRT is minor, extending over less than two points on the whole test.

Commonly suggested means for reducing rater variation include 1) reducing the number of open-ended items; 2) specifying the scoring guide; 3) conducting rater training; and 3) using multiple raters (Meadows & Billington, 2005; Tisi, Whitehouse, Maughan, & Burdett, 2013). The proportion of open-ended items is already low in the NNRT, and it might be argued that lowering it further, or eliminating open-ended items completely, may instead impede on the test's ability to provide a valid representation of authentic reading challenges according to the curriculum (Campbell, 2005, Pearson & Hamm, 2005; Rupp et al., 2006). As noted above, rater training is a complicated measure to take when several thousands of teachers are involved in the scoring process. However, as the administration of test scores in the NNRT is now digitalized, there are technical possibilities for introducing basic systems of multiple ratings of open-ended items and for monitoring rater reliability. Having to adjust one's professional judgement to the judgement of colleagues within the profession would not only be likely to improve fairness for students but also, over time, to reduce the gap between the most severe and the most lenient raters. Such measures will naturally come with additional costs for both administration and teacher scoring time.

On the other hand, if we take into account that 75% of the present item sample is MC, the total scoring reliability of the test is .93. By reducing the gap between the most severe and most lenient raters, the test would be able to include a larger proportion of open-ended items. Suppose the rater reliability of the open-ended items could be raised to .80; then the share of open-ended items could be extended up to 50% of the item sample and total scoring reliability would still be above .90. Yet since students' writing skills is also a component of successful accomplishment of open-ended items, it must be considered to what extent reading test results may vary with students' writing skills.

Another implication from the study concerns the construction of items and scoring guidelines. It is clear that raters disagree about vaguely worded student responses, provided to items where interpretation and reflection is warranted. It should be noted that these responses are problematic from a summative perspective on assessment, but not necessarily from a formative perspective. Vague responses to open-ended items can surely be used as learning opportunities in the classroom, in talking about receiver awareness, in clarifying what it means to interpret text in order to become understood by others, and for helping students to become conscious of their role as test-takers. In the scoring guideline, teachers are encouraged to approve and give credit rather than to fail in cases of grave uncertainty. But test developers must also methodically identify items to which many vague student responses are provided and equip the guideline with a fair number of examples of acceptable and non-acceptable answers. An indication from the study is that a larger number of examples may improve rater reliability.

In addition, it would be recommended that empirical examples of difficult-to-score item responses are used for instance in discussions and meetings for professional development for teachers. In this way, the national test itself may indeed offer more than quantitative estimations of student abilities. It may thus also incite professional dialogue on critical subject-specific issues.

References

Author (2014).

Author & Author, (2016).

Bejar, I. (2012). Rater cognition: Implications for validity. *Educational Measurement: Issues and Practice, 31*(3), 2-9.

Black, P., Harrison, C., Hodgen, J., Marshall, B., & Serret, N. (2011). Can teachers' summative assessments produce dependable results and also enhance classroom learning? *Assessment in Education: Principles, Policy & Practice, 18*(4), 451–469.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch Model* (3rd ed.). New York: Routledge.

Brevik, L. M. (submitted). Assessing reading or assessing readers? Professional development of assessment competence in teachers.

Campbell, J. R. (2005). Single instrument, multiple measures: Considering the use of multiple item formats to assess reading comprehension. In S. G. Paris, & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 347–368). Mahwah, New Jersey: Lawrence Erlbaum Ass.

Cerdan, R., Vidal-Abarca, E., Martinez, T., Gilabert, R., & Gil, L. (2009). Impact of question-answering tasks on search processes and reading comprehension. *Learning and Instruction, 19*(1), 13–27.

Cohen, Jacob (1960). "A coefficient of agreement for nominal scales". *Educational and Psychological Measurement*, *20*(1), 37–46. doi:10.1177/001316446002000104

DeSanti, R. J. & Sullivan, V. G. (1984). Inter-rater reliability of the cloze reading inventory as a qualitative measure of reading comprehension. *Reading Psychology: An International Journal, 5*, 203-208.

EACEA; Eurydice (2009). *National testing of pupils in Europe: objectives, organisation and use of results*. Brussels: Eurydice.

Engelhard, G. (2013). *Invariant Measurement*. New York: Routledge.

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, *61*, 29–48.

Gwet, K. L. (2014). *Handbook of inter-rater reliability*. Gaithersburg, MD: Advanced Analytics, LLC.

Haladyna, T. M. & Rodriguez, M. C. (2013). *Developing and validating test items*. New York: Routledge.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutor Quant Methods Psychol., 8*(1): 23–34.

Illinois State Board of Education (2013). *Illinois standards achievement test 2013. Technical Manual*. Springfield, IL: Illinois State Board of Education, Division of Assessment.

Kane, M. (2013). ”Validating the interpretations and uses of test scores.” *Journal of Educational Measurement, 50*, 1–73.

Kobayashi, M. (2002). Method effects on reading comprehension test performance: text organization and response format. *Language Testing, 19*(2), 193–220.

Koretz, D. (2008). *Measuring up: What educational testing really tells us*. Cambridge, MA: Harvard University Press.

Krippendorff, K. (1980). *Content analysis: An introduction to its methodology*. Beverly Hills, CA: Sage Publications.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* 33, 159–174.

Lie, S., Hopfenbeck, T. N., Ibsen, E., & Turmo, A. (2005). *Nasjonale prøver på ny prøve. Rapport fra en utvalgsundersøkelse for å analysere og vurdere kvaliteten på oppgaver og resultater til nasjonale prøver våren 2005*. Oslo: Institutt for lærerutdanning og skoleutvikling, Universitetet i Oslo.

Linacre, J. M. (2013). *A user’s guide to FACETS. Rasch-model computer programs. Program manual 3.71.0*. Hämtad 2015-04-07. Retrieved from http://www.winsteps.com/a/Facets-ManualPDF.zip

Linacre, J. M. (2014). Facets® (version 3.71.4) [Computer Software]. Beaverton, Oregon: Winsteps.com.

Martin, M. O. & Mullis, I. V. S. (Eds.). (2012). *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

McNamara, T. F. (2000). *Language testing*. Oxford: Oxford University Press.

Mosenthal, P. B. (1996). Understanding the strategies of document literacy and their condition of use. *Journal of Educational Psychology*, *88*(2), 314–332.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part I. *Journal of Applied Measurement*, *4*(4), 386–422.

Kunnskapsdepartementet [KD] (2006, 2013) *Læreplan for grunnskolen og videregående skole* [Curriculum for elementary and secondary school]. Oslo: Kunnskapsdepartementet.

Meadows, M. & Billington, L. (2005). *A review of the literature on marking reliability*. London: National Assessment Agency.

National Directorate of Education and Training (UDIR) (2015). *Nasjonale prøver. Høsten 2015. Vurderingsveiledning lesing 8. og 9 trinn.* Oslo: Utdanningsdirektoratet.

NN (2015). *Den nasjonale prøven i lesing på 8. og 9. trinn, 2015. Rapport basert på populasjonsdata*. Oslo: Institutt for lærerutdanning og skoleforskning. Universitetet i Oslo.

OECD (2009). PISA 2009. Assessment framework: Key competencies in reading, mathematics and science. Retrieved from http://www.oecd.org

OECD (2012). *PISA 2009. Technical Report, PISA*. OECD Publishing. http://dx.doi.org/10.1787/9789264167872-en

Pearson, P. D. & Hamm, D. N. (2005). The assessment of reading comprehension: A review of practices: Past, present, and future. In S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 13-70). Mahwah, NJ: Law. Erlbaum Ass.

Rasch, G. (1980). *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago: The University of Chicago Press.

Rausch, D., & Hartig, J. (2010). Multiple-choice versus open-ended response formats of reading test items: A two-dimensional IRT analysis. *Psychological Test and Assessment Modeling, 52*(4), 354-379.

Rouet, J.-F., Vidal-Abarca, E., Erboul, A. B., & Millogo, V. (2001). Effects of information search tasks on the comprehension of instructional text. *Discourse Processes, 31*(2), 163–186.

Rupp, A.A., Ferne, T., Choi, H., 2006. How assessing reading comprehension with multiple-choice questions shapes the construct: a cognitive processing perspective. *Language Testing 23*, 441–474.

Schumacker, R. E., & Smith, E. V. (2007). Reliability. A Rasch Perspective. *Educational and Psychological Measurement*, *67*(3), 394–409. http://doi.org/10.1177/0013164406294776

Seland, I., Vive, N., & Hovdhaugen, E. (2013). *Evaluering av nasjonale prøver som system.* Rapport 4/2013. Oslo: Nordisk institutt for studier av innovasjon, forskning og utdanning.

Skov, P (2009): Evaluering af brugen af det Nationale kvalitetsvurderingssystem (NKVS) i grundskolen. I: Allerup, P. et al. (2009): Evaluering av det Nasjonale kvalitetsvurderingssystemet for grunnopplæringen. Agderforskning og Danmarks Pædagogiske Universitetsskole ved Aarhus Universitet (s. 103 – 221).

Solheim, O. J. & Skaftun, A. (2009). The problem of semantic openness and constructed response. *Assessment in Education: Principles, Policy & Practice, 16*(2), 149-164.

Taboada, A., Tonks, S. M.,.Wigfield, A. & Guthrie, J.T (2013). Effects of Motivational and Cognitive Variables on Reading Comprehension. In , D. E Alvermann., N. J. Unrau, & R. B. Ruddell (Eds.), *Theoretical models and processes of reading (6th ed.)*. Newark, DE: International Reading Association.

Tisi, J., Whitehouse, G., Maughan, S., & Burdett, N. (2012). *A review of literature on marking reliability research. (Report for Ofqual)*. Slough: NFER.

Wiliam, D. (2013). How is testing supposed to improve schooling? Some reflections. *Measurement: Interdisciplinary Research and Perspectives, 11*(1–2), 55–59.

[1] Note that Kappa was not calculated for these data and that the Kappa value, which takes into account the agreement expected by chance alone, is a more reliable measure and would be lower than the percent agreement measure, especially on short rating scales as in this case (Lie, Hopfenbeck, Ibsen, & Turmo, 2005; Stemler, 2004).