Interrater reliability of constructed response items in standardized tests of reading

Av Michael Tengberg, Astrid Roe og Gustaf B. Skar

Side: 118-137
DOI: 10.18261/issn.1891-5949-2018-02-03
Publisert på Idunn: 2018-05-30

Sammendrag

This article reports from a study of interrater reliability of constructed response items in standardized tests of reading. Two panels of raters (lower secondary teachers and test developers) were asked to rate student responses on 11 different items taken from the Norwegian national reading test in eighth grade. Consensus estimates and measurement estimates were combined with a qualitative analysis of difficult-to-score student responses. Based on findings about rater agreement, distribution of severity, and troublesome response characteristics, the article provides knowledge about both actual and possible levels of interrater reliability and discusses the use and development of open-ended reading test items.

Keywords: Assessment, constructed response, interrater reliability, national tests, reading

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 consistent and scores should be 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 that they propose 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, some 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, their use 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, to provide better empirical grounds for discussing the development of open-ended reading 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 and adjacent agreement extends above 90% (cf. Illinois State Board of Education, 2013). In these cases, large proportions of MC items reassure 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% exact agreement before scoring of the main data collection can begin (Martin & Mullis, 2012). In PISA, there is a similar limit at 85% agreement for any one item and a minimum average agreement at 92% (OECD, 2015). 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 interrater 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 that the scoring of test responses provides (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 are challenging must also 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, for instance, depend on 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 are likely to 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 a 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.

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; Rouet et al., 2001; OECD, 2009).

In a recent pilot study of interrater reliability in the Swedish national reading test in ninth grade, Tengberg & Skar (2016) found that the agreement between raters on open-ended items averaged .73 (Cohen's kappa). A common recommendation is that consensus indicators for interrater reliability should be above .80 (Gwet, 2014), yet such benchmarks must obviously be interpreted in 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 scores 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 (Tengberg & Skar, 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 the University of Oslo and administered by the National Directorate of Education and Training (UDIR) in the autumn term of eighth and ninth grades (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 at school and classroom level. As such it provides educators and school administrators with information about learning needs according to competence aims in the national curriculum (KD, 2006, 2013), for instance through 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 five to seven 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 for each item provides first of all a generic definition of the correct response and of incorrect responses, and then a list of examples of both correct and incorrect responses.

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

After administration, validity and reliability of the test are evaluated in a technical report published online at UDIR's official webpage (cf. Roe, 2014). 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 are 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, it 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:

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?

What characterizes 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 open-ended reading test items.

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 five of the seven texts included in the test and are intended to assess students' ability to retrieve and formulate, in their own words, meaning that 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 eighth grade students (253 responses in all) from an average performing school who had taken the test in 2015. The students in the sample represented different achievement levels and scored between 0 and 11 points on the 11 open-ended items. Responses were distributed to the rater participants 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 (17 women and three men) were lower secondary teachers, who all had several years of experience from scoring national reading tests.

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 seven test developers. 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 compared rater severity and rater reliability in the two groups and used 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 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 seven 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 includes 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 per cent 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 comply with, 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 tests 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). Therefore, it is vital to estimate both actual and possible levels of scoring reliability in the NNRT, in order to gauge viable expectations that may be tied to the test.

Many-facet Rasch measurement

Data was also 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:

$$\mathrm{Ln}(Pni/1–Pni) = Bn – Di,$$

where Pni is the probability of a correct response by person n on item i, Bn is student proficiency for person n, and Di is item difficulty for item i (Bond & Fox, 2015). When Bi = Di, 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:

$$\ln(P_{nij}/1-P_{nij}) = B_n - D_i - C_j,$$

where the added term Cj 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 means that the measures are expressed on an interval scale, which in turn enables the analyst to make relevant comparisons of distances between raters. This can be contrasted to raw score severity measures, which are expressed on an ordinal scale and only allows the analysts to conclude the rank order of raters. By convention, all but one facet is "centred" to have a mean of 0.00 logits.

The FACETS output includes a number of interesting reliability statistics (for technical details see Linacre, 2013; Myford & Wolfe, 2003). For our purposes, the analysis will focus on the distribution of severity among raters over the logit scale, expressed in logits and as separation index, strata ("H-index"), and the reliability of the separation ("R-index"). The H-value can be interpreted as the number of distinct groups of raters in terms of severity. A high R-value indicates that the differences between raters would probably be reproduced in another similar rating. Typically, a test designer would like to have small R- and H-values for raters, indicating more or less inseparable severity levels.

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 analysed in particular the responses for which rater variation was considerably large. The analysis includes 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. Common traits of the difficult-to-score item responses are discussed using examples from those responses (and items) with the lowest rater reliability.

Results

The result section is structured in three parts following the different areas of analysis: kappa statistics of the consensus between raters; many-facet Rasch measurement including reports of separability, fit statistics and rater severity estimates; and, finally, the characteristics of item responses that cause substantial rater variation.

Consensus estimates

In order to investigate the extent of consensus (exact 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 per cent agreement. Figure 1 displays the distribution of agreement between all the 190 pair combinations of teacher raters, showing a distribution from .51 to .89 with a median value of .74.

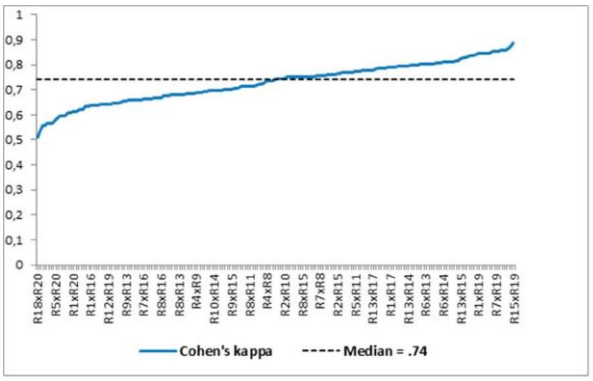

Figure 1. Cohen's kappa for all rater pairs (teachers).

Studying the same statistics for the group of test developers (Figure 2), we can see that the distribution in level of agreement between rater pairs is much narrower. 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.

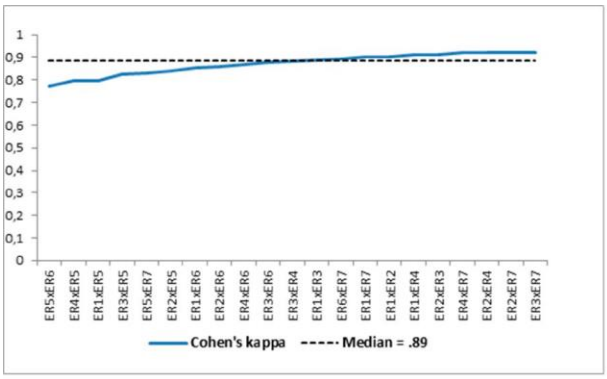

Figure 2. Cohen's kappa for all rater pairs (test developers).

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 and less used in the research literature, is the proper measure of reliability between multiple raters (Gwet, 2008; Landis & Koch, 1977). Fleiss' kappa values 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

Note that because the sample of ratings collected from the test developers is much smaller, the confidence interval for the estimated kappa value is larger, ranging from .84 to .89, whereas for the group of teachers the estimate is much more precise. However, gathering the indications from both the Cohen's and Fleiss' kappa estimates, the results seem to indicate that, for this type of item and format, it is possible to reach a satisfying level of agreement between raters in scoring open-ended responses.

Rasch-modelling

In addition to measures of agreement between raters, the Rasch analysis revealed non-trivial differences in severity.2 As shown in Table 3, the distribution of rater severity over the logit-scale ranged from –0.33 to 0.73. When separating teachers from test developers (see Table 3 and Figures 4 and 5), we notice that the full range of distribution, from –0.33 to 0.73 logits was accounted for by the teacher raters, while the corresponding distribution for the test developers ranged from –0.12 to –0.28 logit. For teacher raters, there was a significant chi-square statistic, meaning that differences between raters were significant. As reported in Table 3, R was .76 and H was 2.70, indicating at least two distinct groups of severity among the teacher raters. For test developers, however, the chi-square statistic was non-significant. Likewise, both R- and H-values indicate non-measurable differences between raters. Thus, in terms of rater severity, the test developers functioned interchangeably.

Table 3.
Severity and separation among raters.

	All raters	**Test dev.**	**Teachers**
Logit min	0.33	–0.28	0.33
Logit max	0.73	–0.12	0.73
Logit mean	0.00	–0.20	0.07
Logit SD	0.30	0.06	0.32
Chi-square	98.5 (26)**	0.8 (6)	81.0 (19)**
R-value	.72	.00	.76
H-value	2.49	0.33	2.70

** $p < .01$

If we look at the raw scores, the most lenient teacher rater awarded 173 points (in average 68%), while the most severe rater awarded 129 points (in average 51%). This difference equalled 1,06 logits or a difference of slightly more than 3 standard deviations, which can be considered to be substantial, representing a 1.87 raw score difference on the test. In sum, the MFRM analysis indicates that there is a relatively unsatisfactory level of interrater agreement among teachers,

which could be attributed to systematic differences in severity. As a contrast, the test developers demonstrate non-significant differences in severity.

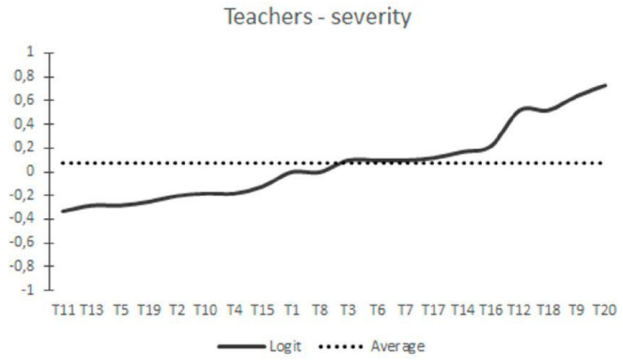

Figure 3. Rater severity among the teachers.

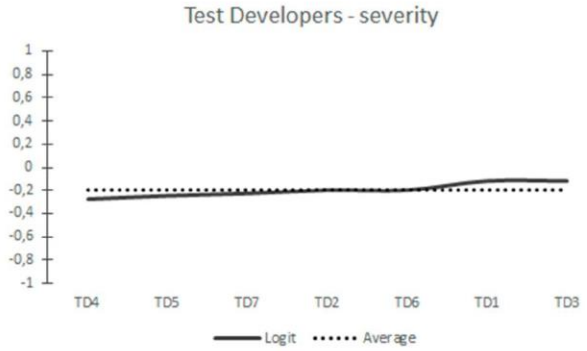

Figure 4. Rater severity among the test developers.

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 raters in the scoring of open-ended responses, we will now describe the qualitative characteristics of responses and items for which agreement between raters was comparatively low. Rating data from the teachers were used to identify problematic items and responses.

Table 4.

Distribution of responses over different levels of agreement.

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

For more than half (52.6%) of the observed responses, there was complete agreement between the 20 teacher participants, and four of five responses (79.4%) obtained 85% agreement or above (see Table 4). However, for 11 of the responses, agreement was 60% or below, meaning that at least eight of the raters disagreed. These 11 responses will serve as empirical examples of responses causing substantial disagreement and are displayed in Table 5 along with item wording and scoring guide.

It should be noted that the distribution of agreement over items can also be attributed to item construction. Three of the open-ended items included in the test produced few or no responses on which the raters disagreed. In these items, the information needed to solve the task is explicit in the text and the scoring guide is unidimensional, thus leaving little room for interpretation on behalf of the rater. In contrast, items that recurrently produce disagreement typically request that the test-taker explains content or form of the text, e.g., to explain a character's emotional state, or why certain information is given 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.

Table 5.

Responses with agreement of 60% or less.

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

Text, item and aspect	**Scoring guide**	**Student response**	**Comment**	Reliability % acc.	
				T[a]	TD[a]
Excerpt from a novel Item: *What surprises the main character at the beginning of the text?* Aspect: 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.	50	86 6 scored 1, 1 scored 0
		What surprised him was that the doorbell rang	Insufficient, lacks an explanation of why the ringing of the doorbell surprised him.	60	86 1 scored 1, 6 scored 0
Excerpt from a novel Item: *In the last sentence it says that the main character was relieved that the skiing trip was over. What may be the reason?* Aspect: reflect	Refers to the main character who finds the skiing trip embarrassing, humiliating for the father, OR who was tired of pretending not to be 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 scored 1, 3 scored 0
Factual text about cartoons Item: *The book got a lot of attention when it first came out in 1954. Why has it become a collector's item?* Aspect: reflect	Refers to the book being a part of 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 them.	The student presents two explanations; one is correct, the other is irrelevant.	55	100 All scored 1
		Because he wrote useful things about cartoons and how they make them.	The response touches vaguely upon a plausible explanation.	60	100 All scored 0
		Because he made horror magazines disappear from the market.	The student indirectly gives a correct explanation.	60	100 All scored 1
Factual text about cartoons Item: *The text is about cartoons, why are different mass media included in the diagrams?* Aspect: 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.	60	71 5 scored 0, 2 scored 1
		To get a better overview of what has gone up and down.	Minimal response that suggests a comparative aspect.	60	100 All scored 1
Factual text about cartoons Item: *Most of the graphs go from 1990 to 2012, but*	Refers to the two graphs and that they represent new media, OR that there	That the thing disappeared or was invented then.	The response provides two explanations, of which one is vaguely	60	86 6 scored 1, 1 scored 0

two of them are shorter, What may be the reason? Aspect: reflect	are no data or information about these media earlier.		related to a correct response.		
		Because it came late.	Short, minimal, implicit answer, Test developers gave the benefit of the doubt.	60	100 All scored 1
Factual text about Hundertwasserhaus Item: *What may be a reason why some people get feel threatened and become aggressive when they visit Hundertwasserhaus?* Aspect: reflect	Refers to a negative view of the house, i.e. that it is ugly, impractical, 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.	60	100 All scored 1

[a] T = Teachers, TD = Test developers

As shown in Table 5, nine of the 11 responses that cause substantial rater variation are connected to reflect-items and two are connected to interpret-items. A common trait of these responses is that they are vaguely worded or insufficient in terms of relevant details from the input text. Thereby, the response itself requires interpretation, through which variation in rater severity will impact the scoring. While some raters will give credit to a response when in doubt, as the scoring guide instructs them to do, others will use vagueness in student responses as an indication of limited comprehension. It is also interesting to note that for types of responses not exemplified in the scoring guide, the teachers gave credit to a lesser degree than the test developers did. This indicates that the number of examples provided may influence the reliability of scoring.

Further, in order to confirm whether vagueness of the student response was typical only for responses causing extensive disagreement, we conducted a thorough review of all the 253 responses. This revealed that although there were a few examples of vaguely worded responses among those where the raters agreed completely, these responses, all of which received credit, typically included information that was easy for the rater to match directly with information in the text or with examples in the scoring guide.

In the following, we analyse in more detail the characteristics of three of the responses included in Table 5, i.e., responses that caused substantial disagreement. One of them was provided to an item related to a narrative text, while the other two were given to an item related to a descriptive text.

Example 1

Item: What surprises the main character at the beginning of the text?

Aspect: Interpret and draw conclusions

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.

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 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. Fifteen 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.

Example 2

Item: The text is about comics, why are different mass media included in the diagrams?

Aspect: Reflect on the content of a text

The item is related to a text about the history of comics, including two line charts which present boys' and girls' use of seven different types of mass media (TV, comics, newspapers etc.) from 1990 to 2012.

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 comics, 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 comics 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 comics earlier. (no comparison with other media)
- To show how comics have decreased and increased. (no comparison with other media)

Two student responses:

1. "To give us more information and that you can read comics on the internet and in the newspaper as well."

(8 teachers scored 1 and 12 teachers scored 0; 2 test developers scored 1 and 5 scored 0).

2. "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 diagram. The scoring guide emphasizes the comparative aspect. The item, however, allows for several different ways of reasoning.

The first response does not highlight a comparative aspect, although it is an indisputable fact that one can read comics 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 by those who gave credit.

The second response can be interpreted as suggesting a comparative aspect by using the word “overview” and by stating that something has gone up, and something else has gone down. The fact that all test developers gave credit to this response may indicate that this implication has been taken into account in the rating. 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 to 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 comics and other media. This item also resulted in disagreement between the raters on 15 of 23 responses.

Discussion

The purpose of the study is to provide more knowledge about actual and possible levels of interrater reliability in the assessment of reading comprehension, and thereby to provide better empirical grounds for discussing the development of open-ended test items. To fulfil this purpose, the NNRT was used as a case for investigating 1) the extent of agreement between both teachers and test developers in the scoring of open-ended items, and 2) the characteristics of item responses for which rater variation was comparatively large. The results of the study indicate that while test developers produced reliable ratings according to both classical test theory measures and MFRM, teacher participants demonstrated significant disagreement, which could be attributed to non-trivial differences of rater severity. Based on closer inspection, we found that the responses that caused substantial disagreement between raters were often vaguely worded and related to interpret and reflect items. Another recurring trait of difficult-to-score item responses was scarcity of relevant details from the input text.

The results of the study 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 action were 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, which 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. Tengberg & Skar, 2016), although both item construction and the structure of scoring guidelines differ substantially between the two tests. In the Swedish case, this level of rater variation may have large impact on students' test results since open-ended items are in the majority and since the items are polytomous, allowing for partial credit scoring. In the NNRT, however, the range of possible variation of student results due to rater variation is minor, extending over less than two points on the whole test. However, while this variation was attributable to systematic differences in severity, further research should investigate whether there may be additional sources of systematic variation. For example, it would be of interest to examine the so-called differential rater functioning (DFR) to detect possible interaction between rater severity and subgroups of test-takers (e.g., good spellers vs. bad spellers). It would also be of interest to examine possible interactions between rater severity and item type (e.g., if raters are more or less severe given an item of a certain type or relating to a certain textual content).

Common suggestions for reducing rater variation include 1) reducing the number of open-ended items; 2) specifying the scoring guide; 3) conducting rater training; and 4) 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 co-rating 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 consists of MC items, the total scoring reliability of the test is still .93.[2] By reducing the gap between the most severe and most lenient raters, the test would thus 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, something that this study has proven to be an attainable level of agreement; then the share of open-ended items could be extended up to 50% of the item sample and total scoring reliability would still be .90.[3] Yet since students' writing skills are also a component of successful accomplishment of open-ended items, it must be considered to what extent reading test results may be allowed to vary with students' writing skills.

Another implication from the study concerns the construction of items and scoring guidelines. The study shows that raters disagreed about vaguely worded student responses

provided to items where interpretation and reflection was requested. These responses are problematic from a summative perspective on assessment, but not necessarily from a formative perspective. In the classroom, vague responses to open-ended questions can be used as learning opportunities when talking about receiver awareness, when 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 is 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.

Appendix A

Table 6. Model fit statistics for all raters.

Rater	Total score	Raw Score Average	Logit	Model	Infit	Infit_z	Outfit	Outfit_z
T5	171	0.68	-0.28	0.16	0.86	-1.70	0.69	-1.70
TD7	168	0.67	-0.23	0.16	0.88	-1.40	0.81	-1.00
TD5	170	0.67	-0.25	0.16	0.88	-1.50	0.75	-1.40
TD2	168	0.66	-0.2	0.16	0.89	-1.30	0.84	-0.80
TD4	171	0.68	-0.28	0.16	0.90	-1.20	0.83	-0.80
T2	168	0.66	-0.2	0.16	0.91	-1.10	0.80	-1.00
TD3	165	0.65	-0.12	0.16	0.92	-1.00	0.84	-0.80
T10	167	0.66	-0.18	0.16	0.92	-1.00	0.91	-0.40

T13	171	0.68	-0.28	0.16	0.93	-0.80	0.93	-0.30
TD6	168	0.66	-0.2	0.16	0.94	-0.70	1.00	0.00
T7	156	0.62	0.1	0.16	0.95	-0.60	1.01	0.10
T4	167	0.66	-0.18	0.16	0.95	-0.60	0.99	0.00
T19	170	0.67	-0.25	0.16	0.95	-0.60	1.02	0.10
T15	165	0.65	-0.12	0.16	0.97	-0.40	0.96	-0.10
T1	160	0.63	0.00	0.16	0.99	0.00	1.12	0.70
T8	160	0.63	0.00	0.16	0.99	0.00	1.05	0.30
T11	173	0.68	-0.33	0.16	0.99	-0.10	1.02	0.10
T17	155	0.61	0.12	0.16	1.00	0.00	1.02	0.10
T6	156	0.62	0.10	0.16	1.00	0.00	1.01	0.00
TD1	165	0.65	-0.12	0.16	1.00	0.00	0.95	-0.20
T3	156	0.62	0.10	0.16	1.01	0.10	1.04	0.20
T14	153	0.60	0.17	0.16	1.03	0.40	1.12	0.70
T18	138	0.55	0.52	0.15	1.05	0.80	1.11	0.60
T20	129	0.51	0.73	0.15	1.14	2.00	1.16	0.90
T16	151	0.60	0.22	0.15	1.17	2.30	1.29	1.60
T9	133	0.53	0.64	0.15	1.18	2.70	1.50	2.60
T12	138	0.55	0.52	0.15	1.22	3.20	1.61	3.20
Average	159.7	0.63	0.00	0.16	0.99	-0.10	1.01	0.10
SD	12.4	0.4	0.30	0.00	0.10	1.30	0.20	1.10

Note: T = Teacher. TD = Test developer. Raters sorted based on infit values.

References

Bejar, I. I. (2012). Rater cognition: Implications for validity. Educational Measurement: Issues and Practice, 31(3), 2–9. https://doi. org/10.1111/j.1745-3992.2012.00238.x

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. https://doi.org/10.1080/0969594x. 2011.557020

Bond, T. G., & Fox, C. M. (2015). Applying the Rasch Model (3rd ed.). New York: Routledge. https://doi.org/10.4324/9781315814698

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. https://doi.org/10.1016/j.learninstruc. 2007.12.003

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37–46. https://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. https://doi.org/10.1080/02702718400 50304

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. https://doi.org/10.4324/9780203073636

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. https://doi.org/10.1348/00071100 6x126600

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. https://doi.org/10.4324/9780203850381

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. Tutorials in Quantitative Methods in Psychology, 8(1), 23–34. https://doi.org/10.20982/tqmp.08.1.p023

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. https://doi.org/10.1111/jedm.12000

Kobayashi, M. (2002). Method effects on reading comprehension test performance: Text organization and response format. Language Testing, 19(2), 193–220. https://doi.org/10.1191/0265532202lt227oa

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. https://doi.org/10.2307/2529310

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. Retrieved 2015-04-07 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.

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.

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

OECD (2015). PISA 2015. Technical Report, PISA. OECD Publishing.

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: Lawrence 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.

Roe, A. (2014): Den nasjonale prøven i lesing på 8. og 9. trinn, 2014. Rapport basert på populasjonsdata. Oslo: Institutt for lærerutdanning og skoleforskning. Universitetet i Oslo.

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. https://doi.org/10. 1207/s15326950dp3102_03

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. https://doi.org/10.1191/026553220 6lt337oa

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. https://doi.org/10.1080/09695940 903075909

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.

Tengberg, M. & Skar, G. B. (2016). Samstämmighet i lärares bedömning av nationella prov i läsförståelse. Nordic Journal of Literacy Research, 2, 1–18.

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. https://doi.org/10.1080/15366367.2013.784165

--

[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).

[2] Since reliability for the other 3/4 of the item sample is 1.0, total scoring reliability can be calculated as (0.72+1+1+1)/4=0.93.

[3] (0.80+0.80+1+1)/4=0.90.