Practical Assessment, Research, and Evaluation

Volume 20 *Volume 20, 2015* Article 9

2015

Interrater reliability in large-scale assessments – Can teachers score national tests reliably without external controls?

Anna Lind Pantzare

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 20, Number 9, April 2015 ISSN 1531-7714

Interrater reliability in large-scale assessments – Can teachers score national tests reliably without external controls?

Anna Lind Pantzare, *Umeå University, Sweden*

In most large-scale assessment systems a set of rather expensive external quality controls are implemented in order to guarantee the quality of interrater reliability. This study empirically examines if teachers' ratings of national tests in mathematics can be reliable without using monitoring, training, or other methods of external quality assurance. A sample of 99 booklets of students' answers to a national test in mathematics was scored by five teachers independently. The interrater reliability was analyzed using consensus and consistency estimates, with the focus on the test as a whole, as well as on individual items. The results show that the estimates are acceptable and in many cases fairly high, irrespective of the reliability measure used. Some plausible explanations for lower interrater reliability in individual items are discussed, and some suggestions are made in the direction of further improving reliability without imposing any system of control.

Teachers are generally trusted to assess and judge their own students for formative purposes, and in some countries also for summative purposes (Harlen, 2005). However, when it comes to scoring external summative large-scale assessments there seems to be a different view. The general procedure when scoring these assessments includes using a set of elaborate and often expensive measures in order to guarantee the reliability and validity of the ratings (Arora, Foy, Mullis, & Martin, 2009; Black, Suto, & Bramley, 2011). The quality of the ratings is secured by central rating with external experts, initial training of judges, and/or monitoring (Baird, Greatorex, & Bell, 2004; Newton, 1996). These commonly used methods to control and ensure interrater reliability are rather complex, and methods such as training and monitoring are often problematic to implement in large-scale assessments where many different judges are involved. However, all of these approaches are, despite the cost, seen as necessary in order to have control over and knowledge about the level of interrater reliability.

This article reports on a study that examines the reliability of teachers' ratings of a national test in mathematics without monitoring, training, or any other method for external quality assurance. Even though the literature recommends the use of different external controls there are assessment systems that can be seen as exceptions. New York State Regents exams are one example where teachers are used as judges. The students' own teachers do not score these exams, but they are marked by other teachers (who are specially trained in marking tests) working at the same school (The University of the State of New York, 2014). Swedish national tests are another exception to the general recommendations made in the literature, and form perhaps an even more decentralized system than the New York example. These tests are scored by the students' own teachers and there are no organized controls in the form of training or monitoring. Therefore it is necessary to investigate if it is naïve to trust the ratings, or if expensive control systems should be implemented.

Methods for improving interrater reliability

There appears to be at least some consensus about possible actions in order to improve the interrater reliability of assessments (Tisi, Whitehouse, Maughan, & Burdett, 2013). First, interrater reliability can be improved by constraining items and closely specifying scoring rubrics (Black et al., 2011). This requires that not only items, but also scoring rubrics, be subjected to field trials. However, constraining items can cause validity problems (e.g. construct underrepresentation), since some competences might require more complex assessments. Second, interrater reliability can be improved by training the judges before the work begins, even if research does point to somewhat different conclusions (see Meadows & Billington, 2005). There are studies that have shown that training and standardization can cause confusion and that the judges can become less willing to use the full range of scores. Despite these results, training is still seen by many as necessary. A third common approach to improve interrater reliability is to monitor ongoing ratings to correct or train the judges if reliability is not satisfactory. Large test-development organizations, such as the College Board in the USA and national agencies in Europe, use trained judges, and they also monitor and moderate the ratings (Black et al., 2011; Newton, 2009; Royal-Dawson & Baird, 2009). Similar approaches are used in international large-scale comparative studies such as TIMSS (Trends in International Mathematics and Science Study). Even though TIMSS has been described as a low-stakes test (i.e. it does not have direct and substantial consequences for the individual student), thorough procedures for high quality rating are still seen as necessary. Judges are subjected to extensive training, interrater reliability is monitored during and after rating, and specific procedures for ensuring quality in rating across countries and across TIMSS "rounds" are employed (Arora et al., 2009). Research has also illuminated the importance of devising methods for controlling rating, for example by using rubrics and moderation to ensure reliability in essay scores (Brown, 2009).

Perfect interrater reliability is only achievable if the rating is fully objective, and such a rating is at least theoretically possible for tests using multiple-choice questions or items that require a very short and closed answer. However, in most cases the judges need to value the qualities of different (and possibly equally relevant) answers to assessment tasks, and thus the risk of differences between judges increases significantly. Black and Curcin (2010) confirmed empirically the expected results that objective assessment items (e.g. multiple-choice, true/false or matching) have the highest levels of interrater agreement followed by short-answer items, and that extended-answer items in general are most difficult to judge consistently. Research has also confirmed that if the judges have scoring rubrics for guidance, the degree of subjectivity in the ratings will reduce (Moskal & Leydens, 2000), and so it is reasonable to assume that the quality of the ratings can be improved significantly by developing effective scoring rubrics (Ahmed & Pollitt, 2011; Brown, 2009; Kane, Crooks, & Cohen, 1999). Bramley (2008) also showed that the larger number of possible scores on the task, the lower the level of interrater agreement.

Studies of interrater reliability in assessment of different school subjects show that assessments in mathematics are most reliable (Murphy, 1978, 1982; Newton, 1996). One reason for this could be the extensive use of multiple-choice and short-answer items. Also, the solutions of many mathematics items follow a definite logic, and this facilitates the development of clear scoring rubrics and thus the likelihood for reliable ratings increases. However, the interrater reliability in these studies is evaluated after training and monitoring the judges. Since many assessment systems have these external quality controls it is not possible to know what the level of interrater reliability would be if the external controls were not in place.

Assessment in Swedish schools

Sweden has a rather unique school system where teachers have a high degree of empowerment and far-reaching responsibilities for teaching, assessing and grading their own pupils (Vedder & O'Dowd, 1999). The teachers grade the students in their own classes without any regular external control. The grading system is criterion-referenced and national syllabi and grading criteria are the basis for teachers' decisions about grades. The grades are used for high-stakes purposes such as educational evaluation in general and for selection to higher education (Wikström, 2005; 2006).

National tests have been developed and used with the primary purpose of supporting teachers in the grading process (The Swedish National Agency for Education, 2005). The teachers judge the national tests for their own students without organized training, monitoring or control. The teachers only have a scoring rubric as a support when rating the tests. Individual student results on the national tests are summarized as a "test-grade" using the grade levels defined in the steering documents, but this result alone does not settle the final course grade for the students. Teachers must also consider other kinds of assessment results (course work, tests, informal assessments, etc.) when deciding on which grade to assign to each student (Dufaux, 2012). Since Swedish national tests have for a long time primarily had the role of supporting teachers' own judgements in relation to grading criteria, it has generally been considered unproblematic to let teachers judge their own students' work without monitoring or other forms of control.

This unique feature of Swedish national tests makes them particularly interesting to study. They are an example of a high-stakes and state-mandated large-scale assessments with low control, built on the assumption that even without control a satisfactory level of interrater reliability can be achieved. Furthermore, the significance of studying interrater reliability in the Swedish context has become more pronounced since the credibility of teacher ratings of student work has been questioned lately. As a result, the Swedish Schools Inspectorate has been commissioned to re-rate a selection of national tests every year (Skolinspektionen, 2011). The results from the re-rating have shown that the agreement between the original rating and the re-rating varies and is sometimes very low, especially in the rating of essays. The methods used in this re-rating procedure are, however, open to criticism (Gustafsson & Erickson, 2013). Indeed, Gustafsson and Erickson conclude that it is not possible to draw any inferences about the quality of the ratings due to flaws in the design of the investigation. For this reason, the need remains to rigorously investigate the quality of teacher ratings, especially in contexts where the teachers are trusted to judge the tests but where there are no external controls.

The overall aim of the study presented here is to empirically examine – from the perspective of interrater reliability – the credibility of teachers' ratings of students' performance on a large-scale assessment in mathematics where there are no external quality controls. The study specifically focusses on the interrater reliability of a Swedish national test in mathematics. These national tests mainly consist of questions demanding an extended answer, and the teachers normally rate the tests only with help of a scoring rubric. The reason for choosing mathematics as the study object is that if rating without training and monitoring does not work in mathematics it would probably not work in other subjects.

The paper is structured as follows. First, there is an elaboration of the theoretical concept of interrater reliability. There then follows method, results and analysis, and finally discussion and conclusions.

Interrater reliability – theoretical underpinnings

Reliability refers in general to the consistency of assessment results of tasks, occasions, judges, groups, etc. (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999).

Whenever students respond to an assessment situation, their performance needs to be evaluated according to some criterion or instruction (in large-scale assessments this is often in the form of a scoring rubric). Interrater reliability refers to the variation in results between different judges evaluating the same student performance, a variation which ideally should be very small.

Interrater reliability has traditionally been treated as a single concept (see e.g. Crocker & Algina, 1986), one that is different from the concept of interrater *agreement*. In this view, interrater reliability is a correlational concept, representing the consistency between judges in the ordering of the performances; in contrast, interrater agreement deals with consistency in absolute terms, representing the degree to which two or more judges make the same judgements on a set of performances (Graham, Milanowski, & Miller, 2012; Tisi et. al., 2013). Stemler (2004) argues that if interrater reliability is treated as a single concept the interpretations of the results could be imprecise, and in the worst case scenario, misleading. Stemler recommends an alternative and more inclusive definition of interrater reliability as representing all types of consistency between judges, including interrater agreement. Instead of only referring to the

concept of interrater reliability one should, when reporting on the quality of the ratings, refer to one of the three proposed general categories: consensus, consistency and measurement estimates. The first category, *consensus estimates*, consists of measures indicating the degree to which two judges agree in the ratings of student performance. These estimates build on the simple notion that judges are expected to interpret and implement the scoring rubric in exactly the same way, and come to the same conclusion when evaluating the same student performances. Exact agreement between two judges is the norm, and the larger the deviances from this state the less reliable the results.

Stemler's second category of measurement for interrater reliability, *consistency estimates*, covers correlational approaches to the problem. These methods build on a view that judges need not necessarily agree as long as they are consistent in their implementation of the scoring rubrics. The use of consistency estimates assumes that interrater reliability allows that judges do not need to share a common understanding of the ratings, but must be consistent in the application of their own interpretation. The severity of judges in their ratings can be accounted for in the process of calculating a student's final score.

The third category, *measurement estimates*, sees variation as an asset and reliability as accomplished by using the information from each judge with respect to an underlying common factor of interest. Interrater reliability is high if only a small amount of measurement error can be attributed to variation between judges. This category builds on the same fundamental view of interrater reliability as consistency estimates, i.e. each judge is expected to be consistent in his or her implementation of the scoring rubrics, but not necessarily in agreement with other judges. Measurement estimates differ from consistency estimates in the sense that they use all information available for each of the judges in order to get a summary score for each test taker. The estimates represent the degree to which scores can be attributed to common ratings rather than errors, and constitute one statistic for multiple judges.

The benefit of referring to these different estimates is that the whole problem of random variation in ratings in relation to judges is captured in the same concept. The consistency, consensus and measurement estimates of interrater reliability are supplementary, and Stemler argues that it might be insufficient to investigate just one of them and that it is important to report which category any statistic represents in order to be able to discuss implications for validity. A consensus estimate can, for example, be high even if a consistency estimate for the same judges is low, and vice versa, implying that the validity judgement can be misleading if only one estimate is used.

The interpretation and application of these categories of estimates raises a question that concerns the relevance of each. Although consistency and measurement estimates have their merits, consensus estimates should be considered as more relevant from a practical point of view. The credibility of large-scale assessments (and any assessment for that matter) is largely based on the expectation that two judges will actually agree in their ratings. We are generally not satisfied by the evidence for the consistency of every judge; we actually want the judges to come to the same conclusions. Furthermore, even though judges can be shown to be consistent in their different implementations of a scoring rubric, results on, for example, Swedish national tests are not corrected for the judges' severity.

Method

Sampling of booklets and judges

Since there is no regular control of interrater reliability in the Swedish national test system, a separate study had to be arranged. In order to conduct a reasonably robust study of interrater reliability, at least two scorers are needed. However, this is often not enough since if the two scorers do not agree, who is correct? Also, if they do agree, can we be sure that they have interpreted the scoring guide correctly? In such cases a third scorer is needed. Adding more scorers can give more information and allow us to generalize from the results. However, with many scorers the analyses become more complicated and the time and cost is increased. Therefore, in this study five upper-secondary school mathematics teachers were commissioned to re-rate a random sample of 99 student booklets of answers to a national test. Initially, 100 student booklets were chosen, but since some tasks were missing in the copy of one of the booklets, this particular booklet had to be removed from the study.

The judges were selected from a group of 15 teachers who had responded to a request to participate in this study. The selection criterions were gender, teaching experience and geography. Two judges were female and three were male. None of the five judges had rated the actual test when it was administered as a regular national test, and therefore the test and scoring rubric were unknown to them before the re-rating. The five teachers came from different parts of the country and did not know each other. They had no contact with each other before, during or after the re-ratings were carried out.

The teachers were experienced classroom practitioners as they all had at least ten years of teaching experience. This corresponds well with the population of teachers involved in the teaching of the mathematics course in question. In the teacher survey connected to the national tests in mathematics, more than 70% of the teachers state that they have this amount of teaching experience. In addition to these five judges, the original ratings in the booklets were also used in the analysis.

In order to simulate the normal conditions for teachers judging Swedish national tests, no training of the judges or discussion about the specific scoring rubric was carried out before the judges started their work. They only received a copy of each answer booklet, a copy of the test and the scoring rubric. The participating teachers were told to score the booklets as they would normally do when scoring national tests. The judges were paid for 50 hours of work to complete the ratings. This time was based on an estimate that it would, on average, take 30 minutes to judge each booklet.

The sample of student booklets used in this study was randomly chosen from among the booklets collected when the national test was administered. Among these collected booklets every fourth booklet was chosen to be included in the study. The booklets were scanned and the original ratings and teacher comments were removed since the presence of the original ratings might have influenced the re-rating (Murphy, 1979). It is not possible to know if the original scoring was made by different teachers since that information is not collected together with the booklets. However, since it is only the booklets from students born on one specific date that are collected, there are normally only one or maybe a couple of booklets collected from each class. Therefore, statistically it might be possible that a few of the booklets were scored by the same original teacher, but it is reasonable to assume that most of the 99 booklets were scored by different teachers.

The test and scoring rubric

The Swedish national test in mathematics used in the study consisted of 16 tasks, comprising 24 items in total, of which one item was multiple-choice (MC), six items were short-answer (SA), and 17 items were extended-answer (EA) where the student had to show all the work leading to the answer. The total test score was 42 and the items had a maximum score of one, two or three points, with the exception of one item that rendered a maximum of six points. In addition to the national test, the teachers always receive a scoring rubric and three cut-scores for the test grades. The cut-scores specifies the number of scores required for each test grade. There are in total four test grades: the first is fail (1) and the other three are passing grades namely, pass (2), pass with distinction (3) and pass with special distinction (4). These test grades correspond to the grading criteria in the syllabi and therefore to the course grades.

The scoring rubric is analytical in the sense that scores are connected to specific parts in the presented solution. The scoring starts at zero and then scores are added when specific parts in the solution are covered. For some of the items there are also, in addition to the scoring rubric, evaluated examples of student work – so-called benchmarks. The purpose of the benchmarks is to clarify and exemplify how the scoring rubric should be interpreted. Depending on the item, the benchmarks can include examples rewarded full score and also partially correct examples.

Statistical methods for calculating interrater reliability

As previously discussed, because of the kind of test analyzed in this study it was thought most appropriate to investigate interrater reliability with consensus estimates. This methodology is also followed in this section of the paper. However, since it might also be important for the judges to be consistent in their ratings the interrater reliability is also analyzed with consistency estimates.

Consensus estimates for interrater reliability can, according to Stemler (2004), be determined by

calculating percent agreement and Cohen's kappa, κ (Cohen, 1960, 1968).

Percent agreement between two judges is the simplest kind of reliability estimate and represents the proportion of students getting the same score when their performance is judged by two different judges. According to Stemler (2004), a rule of thumb is that the percent agreement has to be at least 70%. This can be seen as a very modest demand since TIMSS, for example, has a requirement of at least 85% agreement (Arora et. al., 2009).

One problem with percent agreements is that the statistic can be misleading if most scores fall into one category. This is especially the case when the students have not answered the question, since then there is, of course, nothing to judge. Another problem is that completely random rating will also give a certain degree of agreement. By using κ, the percent agreement is corrected for the amount of agreement that could be expected by chance alone. κ is calculated by:

$$\kappa = \frac{P(a) - P(e)}{1 - P(e)},$$

where $P(a)$ is the observed percent agreement among judges, and $P(e)$ is the expected probability of percent agreement if the rating is done "by chance". $P(e)$ is calculated using the observed data. If the judges are in complete agreement then $\kappa = 1$. If there is no agreement among the judges (other than what would be expected by chance) $\kappa = 0$. Even though κ is widely used in interrater reliability studies there are no absolute levels when κ is defined as being at an acceptable level. Landis and Koch (1977) suggest that values from .61 to .80 are substantial and that values over .81 are almost perfect, a scale referred to in many studies. A disadvantage of the kappa statistic is that it can be somewhat difficult to interpret, since it can differ if the distribution of responses is different for different tasks. Also, the levels of κ seem to be connected to the specific test that is analyzed, hence it is difficult to compare the levels of κ between studies (Sim & Wright, 2005). However, despite these deficiencies κ is often used as one of the measures of consensus.

The statistical methods used to investigate interrater reliability with consistency estimates are correlations and Cronbach's alpha. Correlations are calculated as Pearson correlation coefficients or Spearman's rank coefficients, depending on the characteristics of the data used. A correlation above .70 is seen as acceptable (Stemler, 2004), but in most cases higher correlations between pairs of judges can be expected.

Cronbach's alpha is one of the most commonly used statistics when evaluating reliability of measurements and tests (Cortina, 1993). Alpha can be used as estimating interrater reliability if items are exchanged for judges in the common formula:

$$\alpha = \frac{k}{k-1}\left[1 - \frac{\sum \sigma_k^2}{\sigma_{Total}^2}\right],$$

where k represents the number of judges, $\sum \sigma_k^2$ is the sum of the variances of all judges and σ_{Total}^2 is the variance of the total scores. One advantage of using this statistic is that it yields a single estimate for the consistency of all judges. The rule of thumb for Cronbach's alpha as a measure of interrater reliability is the same as the demands for alpha as a measure of internal consistency, that is, over .70 is acceptable (Nunnally & Bernstein, 1994; Stemler, 2004).

Results and analysis

The data was analyzed with respect to interrater reliability using consensus and consistency estimates, and the results are presented both for students' overall performance on the test (their test grade) and for each item in the test. Cut-scores are based on the total scores and so individual scores are not so important as long as the total makes the cut; from the student perspective it is most important that different judges come to the same conclusion regarding test grades.

Interrater estimates for the whole test

The consensus estimates percent agreement and κ were calculated for all pairs of judges. All agreements were over 80% and several of them were close to or over 90% (see Table 1). In seven of the ten pairs of judges κ was .81 or higher, a level of agreement categorized as almost perfect by Landis and Koch (1977). The rest of the pairs have a kappa of .70–.80, which can be interpreted as a substantial agreement (see Table 1). Crosstabs for all of the pairs of judges are presented in Appendix 1.

Table 1. Consensus estimates for the test grade, above the diagonal percent agreement and below the diagonal κ for the pairs of judges.

	Judge 1	Judge 2	Judge 3	Judge 4	Judge 5
Judge 1	-	94	86	87	91
Judge 2	.91	-	88	93	91
Judge 3	.78	.81	-	91	91
Judge 4	.80	.89	.86	-	90
Judge 5	.86	.86	.86	.84	-

The consistency estimates of judges were determined by calculating pairwise correlations (see Table 2) and Cronbach's alpha. Correlations were calculated both as the Pearson correlation coefficient and Spearman's rank coefficient (Spearman's rho). The results from the two regression methods were similar, but the Table provides the results for the Spearman's rho since the grades cannot be assumed to satisfy the scale requirements for using the Pearson coefficient.

Table 2. Consistency estimates for the test grade, pairwise correlations, Spearman's rho.

	Judge 1	Judge 2	Judge 3	Judge 4	Judge 5
Judge 1	1	.97	.88	.89	.93
Judge 2	-	1	.89	.92	.92
Judge 3	-	-	1	.93	.91
Judge 4	-	-	-	1	.91
Judge 5	-	-	-	-	1

The results show that the pairwise correlations range from .86 to .95; these are fairly high correlations. The second consensus estimate used, Cronbach's alpha, has a value of .98.

Differences between the judges in the study and the original ratings

In addition to the overall analysis of interrater reliability, another aspect of rating differences was specifically studied. In this secondary analysis the ratings passed by the five judges in this study were compared with the original ratings. The mode for the five judges' test grades was compared with the original ratings that had been given by different teachers for all 99 student booklets included in the study.

A system where teachers rate their own students' work can be questioned on the basis that teachers might be biased in favor of their own students. Out of 99 booklets, a difference in the test grade was found in 13 of the booklets (~13%). In ten of these cases the original ratings rendered a higher grade, and in the other three cases the re-ratings were higher. In all but two cases the difference between the original ratings and the mode of the re-ratings was only one point. In nine of the cases the difference was at the cut-score between not pass and pass.

Interrater estimates item by item

Even if the presented consensus and consistency estimates of interrater reliability indicate a fairly high level of overall judgement reliability, the ratings of individual items could still be problematic.

In Table 3, the average percent agreement between all pairs of judges is presented for every item in the test, together with a similar average for κ.

For most of the items, high values for both the average percent agreement and the κ-values are found, indicating substantial or almost perfect agreement. There are, however, a few items with a percent agreement lower than 90% (Items 4, 11a, 13d and 16). An analysis of the student booklets shows that the lower estimates for Items 4 and 13d seem to be due to the fact that it is possible to make an erroneous calculation or use an unacceptable method and still arrive at a correct final answer. Since the answer is correct, some of the judges might not have noticed that an erroneous calculation actually led to the answer given in the scoring rubric. The lower values for Item 11a might be due to an interpretation of the scoring rubric. There were several acceptable answers for this Item, but the scoring rubric gave an example of only one of them. Some of the judges had interpreted this as the only acceptable answer while others had understood that this was one among many acceptable answers. This rather narrow interpretation of the scoring can most likely be found among teachers rating the test, but it might also, to some extent, be an effect of these particular judges' awareness of being part of a study and a resulting tendency to do the ratings more "by the book". Finally, Item 16 rendered a maximum of six scores, and therefore the scoring rubric was inevitably more complicated than for the other polytomous items. The scoring rubric in the national

Table 3. Mean value of pairwise percent agreement and κ between all pairs of judges, for each item. Item types are abbreviated as MC (multiple-choice), SA (short-answer) and EA (extended-answer).

Item number	Item type	Maximum score	Mean percent agreement	Mean κ
1	MC	1	100	.98
2a	SA	1	99	.87
2b	SA	1	98	.91
3a	SA	1	98	.96
3b	SA	1	97	.94
4	EA	3	75	.65
5	EA	3	93	.80
6	SA	1	99	.98
7	EA	2	95	.88
8a	EA	2	93	.90
8b	EA	1	98	.86
9	EA	2	92	.85
10	EA	1	98	.95
11a	EA	2	87	.76
11b	EA	2	92	.86
12	SA	1	96	.90
13a	EA	1	95	.83
13b	EA	1	94	.87
13c	EA	2	92	.85
13d	EA	2	88	.67
14a	EA	1	93	.76
14b	EA	2	95	.84
15	EA	2	93	.78
16	EA	6	78	.68

tests is normally straightforward, with consecutive scores assigned to different stages in the students' work. However, for this Item the order in which crucial steps are taken in solving the problem differs between students, and the order in which partial credits are given can therefore also differ. The scoring rubric can only deal with one of the solution paths explicitly and presents only a more general description of how far the student should have reached in order to be awarded the scores. This could have caused the difference observed between the judges.

From the item-specific analysis it can be concluded that interrater reliability is mainly affected by two factors which are both connected to how strictly the scoring rubric should be interpreted. Firstly, it may be possible to come to a correct answer with an erroneous method or with a different method than that described in the scoring rubric, and secondly, if the student ends with a slightly different answer than that given in the scoring rubric despite using a correct method. In such cases the rating can become complicated. Even if there had been a training session before the ratings were done it is unlikely that these kinds of differences could have entirely been avoided.

Discussion and conclusions

The study empirically examined the degree of interrater reliability for a large-scale assessment when teachers are rating a test consisting of a large number of open-ended items; specific training or other modes of reliability enhancing effort were absent. The study presented here examined the ratings of a test in mathematics, a subject that is usually regarded as rather straightforward in terms of rating. However, this particular test had many complex tasks with more than one aspect to assess and assign credits.

First of all, the results indicate that the overall interrater reliability in this setting is acceptable, and even fairly high according to the recommendations made by Landis and Koch (1977). Also, when compared to other tests where teachers are used as scorers the results are convincing (The University of the State of New York, 2013). This conclusion is based on estimations of interrater reliability from the perspectives described by Stemler (2004) – that is, using consensus and consistency estimates. However, since the study only included five judges the general applicability of the conclusions we reached might be limited. The ability to judge assessments reliably is not inherent, and in a system like the one in place in Sweden, teachers can be expected to develop their ability to read and interpret scoring rubrics from national tests over time. In upper-secondary school, a Swedish teacher will judge national assessments in mathematics every year, or even twice a year. This would support a conclusion that the results in the study could be found rather frequently among experienced teachers. However, in this study there are pairs of judges where the estimates are lower, mainly due to rating mistakes in a few items; Judge 3, for example, did not recognize the erroneous calculation in Item 13d. If those ratings had been corrected the percentage agreement would have risen to nearly 90% instead of the existing level of around 80%.

Second, the suspicion that teachers tend to judge their own students' work more leniently compared to when the students are anonymous was only partially supported by the study. The results show that the original ratings, made by teachers judging their own students' work, were not as lenient as feared compared to the judges for whom each student was anonymous. However, a limitation with this study, and also other studies of interrater reliability, is the effect of actually participating. A relevant question arises from the possibility that the judges were making a greater effort and hence rating more accurately because of their participation in the study. All of the judges knew that they were participating in a study of interrater reliability and therefore they, deliberately or not, might have been more conscious of applying the scoring rubric. They could also be expected to be less lenient in their judgements. However, the only evidence for such a conclusion is that the students who are at the borderline between not passing and passing are, by their own teachers, more often rewarded with the score that is needed in order to pass. The group of teachers who made the original ratings did not know that their work would be scrutinized, and so they probably made the rating as they normally do. In the cases where a difference appeared in the mode of the five judges and the original rating the differences were only one, or in a few cases two, scores. A possible explanation for this may be that the students' own teachers are also grading the students and to do so they will use other information as well as the national tests. If some students showed in earlier coursework that they could reasonably expect to get the grade pass and only one score is missing in the national test, the teacher will probably "find" that score somewhere in the test so they do not have to argue about the test grade with these students.

The third and final conclusion from the study was that the rather small variation between judges can be attributed to difficulties in rating particular items. The analyses identified some characteristic features of items that were causing interrater reliability problems. It seems clear that items where a faulty method can inadvertently lead to a correct answer should be avoided in assessments, a position which also supports the recommendations found in the professional standards literature (see e.g. American Educational Research Association et al., 1999). However, in the literature the focus is often on multiple-choice items where the students are not expected to show their work leading to the final answer, which obviously means that the quality of this work cannot be judged. For the items studied here the students are expected to show their work, and therefore it is possible to see how they have solved the items. Yet despite the presence of this 'solution trail' on paper, there is still a substantial risk that some teachers will not notice that a student has inadvertently produced a correct answer while using incorrect calculations. Such rating errors are identified in this study, contributing to a lowered interrater reliability since some judges do indeed make this mistake and some do not. The level of interrater reliability could have been higher if all accepted methods had been explicitly stated in the scoring rubric. However, if there are many acceptable methods the scoring rubric would be rather complex to read and understand, and that could in turn lead to lowered interrater reliability (Ahmed & Pollitt, 2011).

How can this information about the problematic tasks be used in order to improve the scoring rubric and thereby further improve interrater reliability? For the items where an erroneous method could lead to a correct answer it might be necessary to include such solutions among the benchmarks so that it will be more obvious how to judge them. Also, it is necessary to even more carefully examine the items during the test development process in order to avoid introducing unnecessary difficulties in the ratings.

Despite the support provided by benchmarks, tasks requiring rather extensive solutions seem to result in lower interrater reliability, possibly because of the difficulty in identifying very different solution strategies and assessing their virtues. Excluding tasks of the types described above seems like an obvious way of increasing interrater reliability, but the national curriculum on which these tests are based explicitly states that students are expected to be able to choose from among different methods in mathematics, explain their work and communicate mathematically, etc. Excluding tasks on the sole basis of interrater reliability might therefore have a substantial negative impact on the overall validity of the test.

Even though an acceptable (and even fairly high) interrater reliability has been found in this study, it is not possible to infer that the same results can be achieved in all assessments where teachers are used as independent judges. One important point to consider before any extrapolation can be made is that the task

format makes a difference. The more complex a solution is to judge and the more potentially correct solution strategies there are, the harder it is to achieve high interrater reliability. This is particularly necessary to take into account for complex and extensive tasks, such as essays. Another issue is that although the credibility of interrater reliability studies can plausibly be argued, trustworthiness comes from actually rating observable actions to ensure high reliability: training the judges and monitoring are two such obvious actions. In a system where these kinds of actions are not apparent, this becomes a question of cost and whether it is financially worth the effort or not. As reported, the consistency estimates are rather high but they could always be higher and it is probably possible to improve them with training and monitoring in situations with conditions similar to that presented in this paper. However, it is, from a Swedish perspective, probably not worth investing the money in such a control system in mathematics, at least as long as the national tests are not decisive. Rather, the scoring rubrics should be improved and separate studies to control the interrater reliability can be implemented at regular intervals.

Further research

The presented study only included experienced teachers. The sample was, despite being relatively small, fairly representative of mathematics teachers in Swedish upper-secondary schools who on average have considerable experience of teaching. In order to be able to draw more general conclusions it would be necessary to have a larger group of teachers re-rating tests. Also, it should be borne in mind that in a couple of years the teacher population will change due to retirements, and in order to obtain a more complete picture a sample of less experienced teachers could be included in a similar study. Another extension of the study would be to examine in more detail how the format and structure of the scoring rubric influence interrater reliability, and to see if some sort of linkage exists to items that are particularly difficult to score, i.e. those with several possible answers and pathways to reaching those answers. In particular, a study of how teachers use the benchmarks in the ratings would be interesting. These benchmarks often illuminate how to judge the items when there are several correct methods and answers, and they may reasonably be expected to increase interrater reliability, but in fact very little is known about how they actually function. It would also be interesting to investigate interrater reliability in other subjects, especially those where it is possible to develop similar analytical scoring rubrics as those used in this study.

References

Ahmed, A., & Pollitt, A. (2011). Improving marking quality through a taxonomy of mark schemes. *Assessment in Education: Principles, Policy & Practice, 18*(3), 259-278.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Arora, A., Foy, P., Mullis, I., & Martin, M. (Eds.). (2009). *TIMSS Advanced 2008 Technical report*: TIMSS & PIRLS International Study Center. Boston College, Chestnut Hill, MA.

Baird, J. A., Greatorex, J., & Bell, J. F. (2004). What makes marking reliable? Experiments with UK examinations. *Assessment in Education: Principles, Policy and Practice, 11*(3), 331-348.

Black, B., & Curcin, M. (2010, September 1-4). *Group dynamics in determining 'gold standard' marks for seeding items and subsequent marker agreement.* Paper presented at the annual conference of the British educational research association, University of Warwick, UK.

Black, B., Suto, I., & Bramley, T. (2011). The interrelations of features of questions, mark schemes and examinee responses and their impact upon marker agreement. *Assessment in Education: Principles, Policy & Practice, 18*(3), 295-318.

Bramley, T. (2008, September). *Mark scheme features associated with different levels of marker agreement.* Paper presented at the British Educational Research Association (BERA) annual conference Heriot-Watt university, Edinburgh.

Brown, G. T. L. (2009). The reliability of essay scores: The necessity of rubrics and moderation. In S. D. L. H. Meyer, H. Anderson, R. Fletcher, P. M. Johnston & M. Rees (Ed.), *Tertiary assessment and higher education student outcomes: Policy, practice and research* (pp. 40-48). Wellington, NZ: Ako Aotearoa.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement, 20*(37), 37-46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin, 70*(4), 213-220.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Belmont: Wadsworth Group.

Graham, M., Milanowski, A., & Miller, J. (2012). Measuring and promoting inter-rater agreement of teacher and principal performance ratings. *Washington, DC: Center for Educator Compensation Reform.*

Gustafsson, J.-E., & Erickson, G. (2013). To trust or not to trust?—teacher marking versus external marking of national tests. Educational Assessment, Evaluation and Accountability, 25(1), 69-87.

Harlen, W. (2005). Trusting teachers' judgement: Research evidence of the reliability and validity of teachers' assessment used for summative purposes. *Research Papers in Education, 20*(3), 245-270.

Kane, M. T., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*(2), 5-17.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159.

Meadows, M., & Billington, L. (2005). A review of the literature on marking reliability. *London: Qualifications and Curriculum Authority*.

Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research & Evaluation, 7*(10), 1-11.

Murphy, R. (1978). Reliability of marking in eight GCE examinations. *British Journal of Educational Psychology, 48*(2), 196-200.

Murphy, R. (1979). Removing the marks from examination scripts before re marking them: does it make any difference? *British Journal of Educational Psychology, 49*(1), 73-78.

Murphy, R. (1982). A further report of investigations into the reliability of marking of GCE examinations. *British Journal of Educational Psychology, 52*(1), 58-63.

Newton, P. E. (1996). The reliability of marking of General Certificate of Secondary Education scripts: Mathematics and English. *British Educational Research Journal*, 405-420.

Newton, P. E. (2009). The reliability of results from national curriculum testing in England. *Educational research, 51*(2), 181-212.

Nunnally, J. C., & Bernstein, I. (1994). Psychometric Theory. New York. *NY: McGraw-Hill (1ª ed., 1967).*

Royal-Dawson, L., & Baird, J. A. (2009). Is Teaching Experience Necessary for Reliable Scoring of Extended English Questions? *Educational Measurement: Issues and Practice, 28*(2), 2-8.

Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical therapy*, *85*(3), 257-268.

Skolinspektionen. (2011). Lika eller olika? Omrättning av nationella prov i grundskolan och gymnasieskolan - Redovisning av regeringsuppdrag Dnr. U2009/4877/G. Stockholm: Skolinspektionen.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability [Electronic Version]. *9*(4). Retrieved from http://pareonline.net/getvn.asp?v=9&n=4

The Swedish National Agency for Education. (2005). National Assessment and Grading in the Swedish School System (pp. 32). Stockholm: The Swedish National Agency for Education.

The University of the State of New York. (2013). New York State Regents Examination in Integrated Algebra 2012. Field Test Analysis, Equating Procedure, and Scaling of Operational Test Forms. Technical Report. New York: Prepared for the New York State Education Department by Pearson.

The University of the State of New York. (2014). Regents high school examination. Integrated algebra. Scoring key and rating guide. New York: The state education department.

Tisi, J., Whitehouse, G., Maughan, S., & Burdett, N. (2013). A review of literature on marking reliability research (Report for Ofqual). Slough: NFER.

Vedder, P., & O'Dowd, M. (1999). Swedish primary school pupils' inter-ethnic relationships. *Scandinavian journal of psychology, 40*(3), 221-228.

Wikström, C. (2005). Criterion-referenced measurement for educational evaluation and selection. Umeå: Umeå universitet, Institutionen för beteendevetenskapliga mätningar.

Wikström, C. (2006). Education and assessment in Sweden. *Assessment in Education: Principles, Policy and Practice*, 13(1), 113-128.

.

.

Appendix 1

Pairwise crosstabs for the given test grades for the five judges. In the tables the interpretation of the test grades are 1 = Fail, 2 = Pass, 3 = Pass with distinction and 4 = Pass with special distinction.

Judge 1 * Judge 2

		Judge 2, Test grades				
		1	2	3	4	Total
Judge 1, Test grades	1	30	0	0	0	30
	2	2	45	0	0	47
	3	0	2	11	0	13
	4	0	0	2	7	9
Total		32	47	13	7	99

Judge 1 * Judge 3

		Judge 3, Test grades				
		1	2	3	4	Total
Judge 1, Test grades	1	24	6	0	0	30
	2	3	42	2	0	47
	3	0	2	11	0	13
	4	0	0	1	8	9
Total		27	50	14	8	99

Judge 1 * Judge 4

		Judge 4, Test grades				
		1	2	3	4	Total
Judge 1, Test grades	1	27	3	0	0	30
	2	6	41	0	0	47
	3	0	2	11	0	13
	4	0	0	2	7	9
Total		33	46	13	7	99

Judge 1 * Judge 5

		Judge 5, Test grades				
		1	2	3	4	Total
Judge 1, Test grades	1	28	2	0	0	30
	2	1	45	1	0	47
	3	0	4	9	0	13
	4	0	0	1	8	9
Total		29	51	11	8	99

Judge 2 * Judge 3

		Judge 3, Test grades				
		1	2	3	4	Total
Judge 2, Test grades	1	25	7	0	0	32
	2	2	43	2	0	47
	3	0	0	12	1	13
	4	0	0	0	7	7
Total		27	50	14	8	99

Judge 2 * Judge 4

		Judge 4, Test grades				
		1	2	3	4	Total
Judge 2, Test grades	1	29	3	0	0	32
	2	4	43	0	0	47
	3	0	0	13	0	13
	4	0	0	0	7	7
Total		33	46	13	7	99

Judge 2 * Judge 5

		Judge 5, Test grades				
		1	2	3	4	Total
Judge 2, Test grades	1	28	4	0	0	32
	2	1	45	1	0	47
	3	0	2	10	1	13
	4	0	0	0	7	7
Total		29	51	11	8	99

Judge 3 * Judge 4

		Judge 4, Test grades				
		1	2	3	4	Total
Judge 3, Test grades	1	27	0	0	0	27
	2	6	44	0	0	50
	3	0	2	12	0	14
	4	0	0	1	7	8
Total		33	46	13	7	99

Judge 3 * Judge 5

		Judge 5, Test grades				
		1	2	3	4	Total
Judge 3, Test grades	1	25	2	0	0	27
	2	4	46	0	0	50
	3	0	3	11	0	14
	4	0	0	0	8	8
Total		29	51	11	8	99

Judge 4 * Judge 5

		Judge 5, Test grades				
		1	2	3	4	Total
Judge 4, Test grades	1	28	5	0	0	33
	2	1	44	1	0	46
	3	0	2	10	1	13
	4	0	0	0	7	7
Total		29	51	11	8	99

Citation:

Lind Pantzare, Anna (2015). Interrater reliability in large-scale assessments –
Can teachers score national tests reliably without external controls? *Practical Assessment, Research & Evaluation*, 20(9). Available online: http://pareonline.net/getvn.asp?v=20&n=9

Author:

Anna Lind Pantzare
Department of Applied Educational Science
Umeå Universitet
90187 Umeå Sweden

email: anna.lind.pantzare[at] umu.se