To trust or not to trust? – Contrasting findings on teachers' assessments

Jan-Eric Gustafsson
Gudrun Erickson
Department of Education and Special Education
University of Gothenburg

Trusting teachers' assessments of their own students' performances is a much-debated issue as well as a deep-rooted tradition in many countries. The starting point of the present paper is the Swedish educational system, where teachers are responsible for assigning final grades, essential for students' future choices and opportunities, as well as for marking the national tests that are provided to support their grading decisions. However, this dual responsibility has recently been challenged.

Observations indicating lack of consistency across teachers and schools in the grading of students' performances as well as in marking of the national tests led the Government to mandate the Swedish Schools Inspectorate to remark samples of national tests, and to compare the externally assigned marks with the teacher assigned marks. Two reports from this three-year project have been published, concluding that inter-rater consistency is low and, furthermore, that teachers tend to overestimate their own students' levels of performance. The findings of the Schools Inspectorate have been heavily publicized, leading to distrust in teachers' ratings at the political level as well as among the general public, and among teachers themselves.

In the paper these studies are discussed from a methodological as well as a substantive point of view, focusing on the level of both the individual and the school. One main conclusion is that the actual results from the study are less clearcut than the reports and media releases suggest. Results from the Schools Inspectorate's studies are compared to the findings of other recent studies that have also investigated teachers' markings of national tests. In the discussion the results and implications of trust versus distrust in teachers' assessments are considered from validity-related as well as from ethical points of view.

Context

The current paper is set in the Swedish educational context, which, on the one hand can be characterized as clearly centralized, with national curricula and syllabi for individual subjects, and a national Schools Inspectorate to control that these documents are correctly used, on the other hand as distinctly decentralized, relying heavily on what is often referred to as 'teachers' professionalism', i.e. on decisions taken by individual teachers. These decisions range from content and methodology to assessment and grading. Since grades are commonly used for selection for higher education, this requires a high degree of assessment literacy among teachers.

There is a long tradition of national assessment in Sweden, comprising formative as well as summative materials and tests, the latter mandatory for schools to use and commonly perceived as high-stakes. The national tests are aimed to supplement and support teachers' grading, not to function as exams in the traditional sense. There are, however, no regulations or guidance concerning the weight of the national test results in these decisions; individual teachers decide on individual students' grades. Moreover, there is no central marking, and, in the vast majority of all cases, teachers mark their own students' tests.

The national tests are developed by different universities in the country, commissioned by the National Agency for Education (NAE). Test development is done in collaboration with a number of stakeholder categories, e.g. teachers, teacher educators, researchers, and students. The opinions of the latter are systematically collected during piloting phases, and the data are analysed and incorporated in the decision process preceding a live test (Erickson and Åberg-Bengtsson, *forthcoming*). The national tests are generally well received by teachers (> 90 % positive judgements). Results are analysed and published in annual, publically available reports on the NAE website.

The national tests comprise extensive guidelines for teachers, touching on issues of purpose, construct, rubric and use. In addition, comments are given on different types of responses and ratings, and a large number of benchmarks are presented to strengthen the accuracy and consistency of teachers' markings. Successive studies are undertaken to analyse the outcomes of the tests. In this, validity, reliability and functionality for purpose are important aspects, as will be further commented on later in this paper.

Background

As has already been mentioned, observations indicating lack of consistency across teachers and schools in the grading of students' performances as well as in the marking of the national tests led the Government to mandate the Swedish Schools Inspectorate (SSI) to remark samples of national tests, and to compare the externally assigned marks with the teacher assigned marks. The Government also provides a background where theoretical and practical starting points are discussed.

A study by the Swedish National Agency for Education (NAE) (2008) investigated issues in moving from teacher marking of the national tests to central marking . One main conclusion was that this would increase the yearly cost for the national tests at national level from around 37 mSEK to around 200 mSEK. The NAE (2008) also reviewed the available Swedish research on reliability issues in the marking of national tests. Very few systematic large-scale Swedish studies have been conducted, but the university projects developing the national tests have reported several small-scale studies. For mathematics three studies were discussed, which all found a small tendency for the individual teacher to give a higher mark than was assigned by independent markers. In the field of English

one study is referred to, and in this too a very small tendency was found for the teachers to be more lenient in their marking than independent markers. This effect was in particular seen for a small set of items which only had general instructions for the marking. In Swedish two studies are referred to, one of which showed relatively large differences between marks assigned by teachers and independent raters, the teachers being more lenient. The other study did not demonstrate any such systematic difference, even though it showed that there were quite considerable inconsistencies among the marks assigned to the same student by different raters. Low interrater reliability has also been found in a study of an oral subtest in the national test of mathematics C in upper secondary school.

One main conclusion from the NAE (2008) study was that the results do not indicate that teachers' marking of national tests is a major source of bias, even though there is a tendency for teachers to assign higher marks than exterrnal markers. Another main conclusion was that there may be threats to the trust in the system of national tests by important stakeholders, if the system of teachers' marking comes to be regarded as unreliable. However, the high level of cost for introducing central marking of all national tests was regarded as prohibitive, and instead a series of other measures were suggested to improve reliability and credibility of teachers' markings, such as exchanging tests between teachers; increased collaboration in the marking between teachers and schools; sample based quality control of the teachers' marking; and instructions to the SSI to make inspection of marking of national tests a part of the regular inspection. The NAE (2008) also suggested that systematic research should be conducted on issues of marking of national tests.

The Government referred to the NAE (2008) report and concluded that there is no dependable basis for assessing whether the national tests are marked correctly. However, the government also stated that the degree of agreement between the final grades assigned by the teachers and the test marks is only about 80 % and also that there are considerable differences in the amount of discrepancy between schools, remaining stable over time. It was suggested that this pattern of results may be indicative of suffering comparability between schools in the marking of the national tests and in the grading of students.

In the preparatory work for the establishment of the SSI in 2008 (SOU 2007:101) it was suggested that it could include a function to secure the quality of grading of samples of national tests and to compare this with the grading in the different subjects. The major function would thus be to detect incorrect markings which jeopardize the reliability of the system of national tests. In the decision made by the government this suggestion was followed, and it was decided that samples of at most 10 % of the national tests should be made each year for three years, at a maximum cost of 20 mSEK per year.

The Government's assignment to the Swedish Schools Inspectorate

The SSI was instructed to collect a sample of copies of already marked student papers and to remark these. It is also stated that the remarking of the student responses is to be a part of the running inspection activities of the SSI, and that the results of the remarking can be made a basis for conducting deepened inspection activities within the regular inspections and thematic quality inspections.

According to the instructions the remarking is to be made in three rounds, to be conducted during three successive years. In the first round a sample is to be made among the obligatory national tests in compulsory school and in upper secondary school. The sampling shall be done in such a way that the results are representative at national level for each subject and grade, and also at school level among the selected schools in Swedish, English and mathematics in grade 9 and in upper secondary school. The sampling shall also cover national tests in new subjects which are under development.

The SSI also was instructed to analyze the outcome of the remarking of a representative sample of student responses with respect to the presence of systematic misjudgement. To the extent that the SSI concludes that there is systematic bias in the marking of the tests, the SSI is instructed to conduct a closer analysis of these. If the results indicate need for a closer inspection, the SSI is to conduct the inspection on a new sample of schools. Yet another task for the SSI was to suggest suitable grounds for selection in the next two rounds of remarking.

The implementation and results of the remarkings in 2010

The report concerning the remarking of tests taken in the spring of 2009 was delivered to the Government in April 2010 (Skolinspektionen, 2010). It was based on remarking of some 29 000 national tests from 633 schools. The national tests were Swedish, mathematics and English in grades 3 (not English), 5, 9 and the first obligatory course in the same subjects in upper secondary school.

The results were reported in terms of deviations between the original teacher marking and the remarking. The reporting was done at school level, and the number of negative (i.e., original mark higher), positive (i.e., original mark lower) and agreeing marks were tabulated for each school.

The sample of tests was not fully representative of the population, and there were several reasons for this. All schools that were asked to send copies of the student responses did not comply, and some sent the material too late, or only for subsets of the students. In some cases the copies were too difficult to read to be included in the study. On the basis of analyses conducted by Statistics Sweden of the representativeness of the obtained sample of test responses, it

was, however, concluded that the sample was representative at national level for most subjects.

The major finding was that the remarking resulted in substantial differences between the original teacher mark and the control mark for those subtests where the student is to produce a longer text in Swedish and English in grade 9, and in Swedish, English and Mathematics for upper secondary school. The same tendency, but not equally pronounced, was observed for grades 3 and 5.

The observed differences were both positive and negative, but negative differences were much more common, and particularly so for the subtests in grade 9 and upper secondary school. It was also observed that the negative differences were larger in the upper part of the scale of marks, and it was in particular observed that the control marking tended to differ from the original marking at the highest mark.

The SSI concluded that it is not possible to determine whether it is the original mark or the control mark that is correct, but they nevertheless emphasized that the fact that the outcome depends so much on which teacher does the marking is worrying from a comparability persective. The SSI also concluded that the results throw doubts on whether the national tests are constructed in such a way that they can fulfill their main aim of supporting grading which is comparable and fair. They identified three main problems in the construction of the tests that may explain the discrepances found.

The first problem is that, at the time of the first data collection, the national tests had a wide range of aims: In addition to providing support for teachers in their grading of students' competences, the tests were aimed to increase goal fulfillment, concretize goals and grading criteria, and provide a basis for evaluation at individual, school, municipal and national levels. The SSI emphasized that it may not be possible to reach all these aims with one and the same test.

A second reason for the poor agreement noticed between the markings, as suggested by the SSI, was that the tests and the marking instructions leave large margins for interpretation, which does not support comparable judgements. This applies in particular to tasks in which the students are to provide their responses through writing a longer text.

The third reason for the tendency for the teachers to be more lenient than the control markers, suggested by the SSI, is that they are less objective. The SSI refers to anecdotal evidence from teachers hired as control markers, saying that they interpret the marking instructions more strictly when they are control markers than when they mark their own students' tests, because in the former case they do not have any relation to the student.

The SSI also concludes that the differences between the original marks and the control marks vary between schools, with few differences for some schools and substantial differences for other schools. In the report this phenomenon is not analyzed further, but it is stated that the SSI will follow up the marking in schools with large differences.

The report ends with a set of recommendations: (1) the primary purpose of the tests should be clarified; (2) the tests should be constructed in such a way that the primary purpose can be reached ; and (3) a marking system where the identity of the student marked is not known should be introduced.

The implementation and results of the remarkings in 2011

Another conclusion of the 2010 report was that the sampling for the next round of remarking should be conducted in the same way as the first round. The report concerning the remarking of tests taken in the spring of 2010 was delivered to the government in April 2011 (Skolinspektionen, 2011). It was based on the remarking of some 35 000 national tests from 750 schools, and the procedures followed were the same as those applied the previous year.

The results also were more or less identical. The report emphasizes that there were substantial differences between the original marking and the remarking for certain subtests, and particularly for those subtests in which the student is to respond in the form of an essay. It is thus observed that there were differences in about 50 % of the tests in Swedish in upper secondary school. In about 20 % of the cases the marks differed by more than one step on the four point scale. The differences were, furthermore, in most cases negative, meaning that the SSI assigned a lower mark than did the teacher.

The report also concluded that there were large differences between schools in the amount of deviation between the two markings. Here too, however, the SSI concludes that the study does not make it possible to determine whether discrepances observed for single schools are systematic or not.

The three main conclusions from the first report were repeated in the second report.

Reflections on the design and conclusions of the SSI studies

The main conclusion drawn by SSI is that the mark assigned to a student's subtest to a large extent is connected with the individual marker, and the SSI therefore questions whether the national tests can be claimed to support comparable marking and grading.

There are, however, two problems with this conclusion. The first problem concerns the conclusion that there are differences between teachers in which

marking standards they apply. It is indeed a well-established fact that comes out of almost any study of marker characteristics that some are more lenient ("doves") while others are more harsh ("hawks"). But if this holds true for the general population of teachers, it is highly plausible that there is also variation in leniency among the teachers recruited by the SSI to do the remarking.

This raises several questions. The first is if the sample of teachers employed by the SSI is representative of the whole population of teachers. The 2010 report states that a private company recruited 211 teachers to conduct the remarking (178 teachers in the 2011 report). These teachers were to have a teacher education adequate for the subject and grade, as well as experience in marking tests in the subject and grade that they were to remark. They did the remarking in their homes but were supported by SSI personnel through digital communication. While the two SSI reports present careful analyses of the representativeness of the samples of students and schools, no attention whatsoever is directed to the characteristics of the sample of teacher for the remarking. Given that this sample is more or less self-selected, it is not likely to be representative of the general population of teachers in the investigated subjects and grades, but it is, of course, impossible to tell in what respects they differ from the population.

Another important question is how the remarking teachers were matched with the marking teachers and the tests. This is not clearly described in the report, but according to information from the SSI, the remarking teacher remarked all the subtests of a particular national test. Also, a remarking teacher was allowed to remark a maximum of 10 tests from a particular school, in order not to allow too strong an impact of a particular remarking teacher on the results for a certain school. This information indicates that the SSI is at least vaguely aware that the selection of remarking teachers might influence the results. However, it seems that they did not systematically design the study to take possible variability among the remarking teachers into account.

From a design perspective it would have been optimal to use as many remarking teachers as possible for the students from each school. In the first study there were, for example, 34 schools in the sample of English national tests in grade 9, with an average of 70 students in each school (range 29 to 128). With the large number of remarking teachers available, this would have made it possible to assign zero, one or two tests from each school to each remarking teacher. With such a design it would have been possible to get good estimates both of variation in leniency among the remarking teachers and of the variation in leniency in the original marking of the tests among the schools.

It may also be noted that different samples of schools were selected for the different national tests. Such a sampling design has the advantage that the amount of work for each sampled school is smaller than if the school were to copy the results for all the national tests for each student. But it is not obvious that such a sampling design is optimal, given that the purpose is to study

variation among teachers and schools in the marking of national tests. If each school is represented by one subject matter only, there is no basis for investigating whether school differences are consistent over subjects. Furthermore, given that each school is represented by the few teachers who teach a particular subject matter, school differences will to a large extent be confounded with teacher differences. Another possibility would have been to include tests from all subjects from each school, but not necessarily include all students within each school. Such a design would allow inferences about between- and within-school differences in the marking of national tests, information that seems essential to be able to interpret the findings.

The second problem with the SSI conclusion that the national tests do not support comparable marking and grading, is that this conclusion is based on results from analyses of components of the full test. However, the support for grading is typically assumed to be provided by the results on the full tests, and not by the results on a single component. Thus, the conclusion drawn by the SSI may be correct or it may be incorrect, because it does not follow from the results presented. Should it be that the test component which is identified to be unreliable constitutes a major part of the total test result, the conclusion drawn is more likely to be correct, but if the test component only has a minor influence on the total test score, the conclusion is more likely to be incorrect. Given that there are considerable differences in the design of the different national tests, the effects of unreliability of scoring must be evaluated for each national test separately, but this has not been done by the SSI.

The data collected by the SSI should be reanalyzed with a focus on differences between different remarking teachers and on differences in outcomes for different subtests within and between schools. Currently we do not have access to data, which allow the first type of analyses, but from the information in the published reports we can take a closer look at the second question.

A closer look at outcomes for different subtests

We focus on the results for the grade 9 tests in English and Swedish. One reason for this is that in the assignment to the SSI special emphasis is put on the national tests given at the end of comprehensive school, and in particular in the analysis of school differences. Another reason is that for these two tests, results are presented for two subtests each, which makes it possible to investigate degree of consistency of school differences in amount of deviations between the original marking and the remarking. For the national test in mathematics only a global result is presented, so data for this test it is not useful for the current purpose.

The national test of English for grade 9 consists of three parts, measuring oral interaction and production (Part A), receptive skills when reading or listening (Part B), and written production (Part C). There is no documentation of student

performance on Part A (recording is recommended but not obligatory), so it was not included in the remarking, but Parts B and C were. Part B consists of two sections, listening comprehension (in total c. 35-40 items) and reading comprehension (c. 55-60 items), the number of items per skill varying marginally from one year to the other. In this part the students respond by writing shorter or longer answers or by fixed response options (multiple choice, matching). In Part C the students write a text, choosing from two alternative topics, one which allows more freedom and one which is more structured.

The national test of Swedish consists of three subtests. Subtest A tests reading comprehension of different types of texts, and the student's ability to express his/her own reflections about texts. Subtest B is conducted in groups and tests the ability to communicate clearly and comprehensibly. This subtest is not included in the remarking. Subtest C is a written essay, which tests the ability to develop and express thoughts and ideas, and to produce narrative and descriptive text. This subtest is included in the remarking, which thus comprises subtests A and C.

The SSI reports present tables in which the results for each school and subtest included in the remarking are reported: the number of negative differences (original mark higher), positive differences (original mark lower) and the number of marks in agreement. The information in these tables has been the starting point for computing the difference between the percentage of positive differences and the percentage of negative differences. These net differences thus express the balance between the number of original marks lower than external marks and the number of original marks higher than external marks, negative numbers expressing lower marks in the remarking, and positive numbers expressing higher marks in the remarking.

Table 1. Statistics for differences between original marking and remarking at school level

Subtest	N	Minimum	Maximum	Mean	Std. Deviation	Correlation
English B 2010	34	-12.9	14.0	-1.9	6.2	
English C 2010	34	-55.5	25.4	-13.4	17.7	0.24
English B 2011	43	-14.8	10.5	-3.2	5.4	
English C 2011	43	-50.0	34.3	-12.2	16.7	0.26
Swedish A 2010	36	-48.8	15.0	-19.9	14.4	
Swedish C 2010	36	-64.3	27.8	-21.0	18.0	0.37
Swedish A 2011	40	-58.8	20.0	-24.0	16.9	
Swedish C 2011	40	-66.1	19.2	-24.2	18.2	0.58

Table 1 presents descriptive statistics for these percentage differences for English B and C and for Swedish A and C for the two rounds of remarking. The mean values are all negative, indicating that the marks assigned by the teachers tend to be higher than the marks assigned by the SSI. There are, however, substantial differences between the different tests and subtests in the magnitude of differences. For English Part B the means are close to zero, while for English

Part C they are around -13 %, indicating a small tendency for higher teacher than SSI marks. For both the Swedish subtests the means are around -20 % in the first round and -24 % in the second round, indicating that marks assigned by the teacher are substantially higher than marks assigned by the SSI.

The standard deviations express the variation in degree of imbalance at school level, as do the minimum and maximum values. The standard deviations are substantial, indicating that the original marking in some schools is much more lenient than the SSI marking, while in other schools the original marking is less lenient than the SSI marling. The minimum and maximum values provide the same impression in a more concrete manner. It is interesting to note that the standard deviations are quite similar for most subtests with values around 17. The English B subtest forms an exception, though, with standard deviations around 6.

Table 1 also presents correlations between subtests within each national test. If schools have balances of mark differences in the same direction for the subtests, correlations will be high, but if the balances do not go in the same direction for the schools the correlations will tend towards zero. For English the correlations are low (around .25) and non-significant. For Swedish the correlations are somewhat higher and particularly so for the second round (.37 and .58 for the two rounds, respectively) and both are statistically significant. However, even the highest correlation implies that only around 34 % of the variance in the two subtests is common.

Discussion and conclusions

Below we discuss the results and interpretations of the empirical results, we comment on the way the SSI study has been conceived and conducted, and we reflect on the issue of trust.

Reliability and bias in marking of tests

The results presented above are in agreement with the results presented by the SSI in showing that subtests involving production of longer texts are assigned higher marks by the external markers than by the students' teachers. However, our results, somewhat unexpectedly, show that there is little consistency in school differences between external and internal marking for different subtests. This is clearly the case for English while for Swedish there is low positive correlation in the first round of remarking, and a higher correlation at the second round.

The result that marking of longer essays tends to be unreliable is a well-established fact, many studies indicating that inter-rater reliabilities do not exceed .50. A recent study conducted in collaboration between the Swedish NAE and the university departments responsible for the development of the national tests

investigated inter-rater consistency in different national tests and provides interesting results for comparison (Skolverket, 2009). The study included, among others, English, mathematics and Swedish in grade 9, i.e. at the end of compulsory school. 100 randomly selected, teacher-marked tests were analyzed and rated by two or three independent raters, two for closed formats with dichotomous rating, three for tasks requiring constructed response of varying length and complexity. The results (Skolverket, 2009; e.g., Erickson, 2009) showed high inter-rater correlations for English and mathematics (between .86 and .99 for English, and between .81 and .99 in mathematics), whereas the inter-rater reliabilities were distinctly lower for Swedish (between .36 and .86).

While reliabilities around .50 is lower than is normally required in measurement situations it does seem to be a reflection of the inherent difficulty of the task to assign a complex and multifaceted piece of work into a small set of categories representing different marks. It should in this context also be observed that the remarking conducted by the SSI was done on a scale with four categories (i.e., the four marks Fail, Pass, Pass with Distinction and Pass with Special Distinction), while the teachers' original marking of the national tests was done on a ten-point scale. In this scale a distinction is made between two levels (weaker and stronger) for the Fail and Pass with special distinction levels, and three for the Pass and Pass with Distinction levels. Given that classification into fewer categories always yields a lower inter-marker reliability than classification into a larger number of categories, this provides another explanation for the low observed reliability in the SSI study. (As an example, it could be mentioned that using a four-point scale in the study of the national test of English referred to above would decrease the correlations for Part C /Writing/ by about .10, from .86-.93 to .78-.82.)

Let us also add that even though reliabilities around .50 are unacceptably low in a high-stakes situation, there are other high-stakes situations where reliabilities are even lower. For example, Marsh (e.g., Jayasinghe, Marsh, & Bond, 2001) has shown in a series of studies that the inter-rater reliability of peer-reviews of proposals for research funding is around .20. Such low reliabilities necessitate use of multiple raters, and in well-designed systems for evaluation of research proposals a sufficient number of reviewers is used to achieve satisfactory reliability. While only a single marker (i.e., the teacher) is used in the Swedish grading system, the final grade not only depends on the results obtained on the national test, but also on the results obtained on other written composition tasks conducted over several years, and also, of course, on assessment of other goals than written composition. Thus, even though the results show that the reliability of marking a single essay is low, the fact that the grade is based on assessment of performance on multiple tasks indicate a higher reliability in the teacher assigned grade.

However, the result most heavily emphasized by the SSI is not the low inter-marker reliability but the finding that the marks assigned by the external markers

tend to be lower than those assigned by the teachers. While the published reports do not make strong claims that the teachers are too lenient in their marking, this is nevertheless the message that has been conveyed by the SSI to media.

As mentioned earlier, several small-scale studies on inter-rater consistency have been conducted by the national testing teams at different universities in Sweden as part of the regular validation successively undertaken. Results have shown a weak tendency of teacher leniency in marking their own students' tests. The same tendency has been demonstrated in other contexts and countries, as reported for example by McKinstry et al. (2004) and Harlen (2005). In the study by Skolverket (2009) teachers' ratings were found to be somewhat higher than those by external raters, however almost negligible in English and mathematics, and distinctly more marked in Swedish.

There are several different interpretations of the difference in leniency between internal and external marking. One is, of course, the one that, at least implicitly, seems to be favored by the SSI, namely that teachers are positively biased in the assessments of their own students. Another possible interpretation is that the external raters are negatively biased in their assessments, because they have interpreted their remarking assignment in such a may that they become harsher in their assessments. The anecdotal evidence reported by the SSI provides some support for this interpretation.

Other interpretations also are possible. The fact that the external markers used a four-point scale while the original teacher ratings were done on a ten-point scale causes underestimation of inter-marker reliability and it may also have affected leniency. The re-marking was done on the basis of hand-written copies of the students' papers, some of which were more or less illegible. This may have affected the possibilities to interpret the text as intended, thereby lowering the mark. Along a similar line of reasoning it may be hypothesized that familiarity with a student's handwriting and modes of expression makes it easier to interpret the intended meaning, leading to a more positive evaluation.

The main conclusion that can be drawn is that the SSI remarking provides strong evidence that marking of longer written texts by an external examiner results in lower marks than when marked by the teacher. This finding is in agreement with much other research, but it is not clear how the finding can be explained.

Should it be that the amount of difference between external and internal marking is constant for all students and schools, this does not influence the rank-ordering among students and schools and in such a case it does not matter whether external or internal marking is used. However, if the amount of difference in the marking of the national tests varies over schools, such that teachers at certain schools are more lenient, and others are less lenient, this may have important consequences both for schools and for students, given that the final grades are

regarded as important sources of information about the quality both of schools and students.

The results clearly show that there is considerable variation among schools in leniency of marking of longer written texts. However, the results also indicate that these differences tend to be inconsistent across different subtests within each national test, even though there is more of consistency for Swedish than for English. If the differences are inconsistent across subtests this lessens their influence on the mark for the full national test, because unrelated differences will tend to cancel. Furthermore, if the differences are inconsistent across the three national tests set in grade 9 there will be less of a systematic effect of leniency than if the differences at school level go in the same direction. Regrettably, however, the SSI has chosen to design the study in such a way that we only know little about degree of consistency over subtests at school level, and nothing at all about degree of consistency across the different national tests.

To the extent that teachers differ in degree of leniency in their marking this will cause some students to earn higher marks on the national test than other students with the same level of knowledge and skill. This is a source of unreliability that may be regarded as a problem of fairness in comparisons among individual students. Because there typically are few teachers within each school who mark tests within each subject, teacher differences also influence school differences. Again, however, we have to regret that the design used by the SSI is uninformative with respect to differences in leniency both among the teachers who did the original marking of tests, and the teachers who did the re-marking.

The logic of the SSI study

The SSI was given the assignment, by the Government, to investigate if there is bias when teachers mark their own students' tests. This is a seemingly simple question, of a kind that is being approached in research on educational measurement. However, it is easy to see that the SSI did not adopt the methods and theories within the field of educational measurement. Instead they adopted theories and methods from the field of school inspections when they tried to answer the question. Given that the SSI obviously have failed to produce any useful answer to the question, at the same time as they have had an immense impact on the trust of the general public in teachers' ability to assign unbiased marks, it is of interest to discuss the underlying logic of the SSI study.

In the field of school inspections a set of beliefs and theories have been developed about how inspection activities improve practices of schools, and methods and procedures have been developed in order to collect relevant information (Gustafsson & Myrberg, 2011). These beliefs, theories and methods differ in fundamental ways from the beliefs, theories and methods used in research, and below we make some comments on these differences.

A first difference is that while in research the participants are typically anonymous, this is not the case in school inspections. Thus, while researchers typically report aggregated statistical results in tables and graphs, the tables and graphs in the SSI reports present identifiable schools. This is likely to be due to the fact that school inspections use a “naming and shaming” approach to correct deviations from rules and regulations, both for the individual school and for the general impact on other schools (Gustafsson & Myrberg, 2011). Such an approach essentially assumes that knowledge is available about the correct way to do things, and the task of school inspection is to identify schools that do things in incorrect ways. Thus, it is not an aim of school inspection to search for new knowledge but to identify violations of rules and regulations, and correct these. For the re-marking project the SSI adopted a typical inspection design, which may be useful for purposes of school inspection, but which is useless for purposes of developing knowledge about bias in the marking of tests.

A second difference is that while in research much attention is focused on characteristics of the instruments used for data collection, such as reliability and validity, in school inspections it is taken for granted that inspectors have the capacity to make the correct observations and interpretations. One reason for this is that research is an open activity in which the research community has the task to carefully scrutinize procedures, findings, and inferences in all kinds of research. School inspections are typically not challenged and scrutinized by fellow inspectors, and even though it has occasionally been observed that there is a need to attend to issues of validity and reliability (e.g. SOU 2007:11), this is rarely done. This may explain why the SSI in their reports focus all their attention on characteristics of the schools, the samples of tests and the teachers who contributed the original marks, but do not provide any information whatsoever about the teachers who did the re-marking, because as representatives of the SSI their marks must be both valid and reliable.

A third difference is that while reports of research typically go through a long process of review before eventual publication, the reports from the SSI go directly to media, where they often result in front-page headlines. This causes the reports to have high impact, and there are many indications that the reports on bias in teachers’ marking of national tests have caused loss of trust in teachers. Furthermore, while reports of research are often discussed at length in groups of specialists, there is little opportunity to discuss inspection reports, neither among specialists, nor among the general public.

In conclusion, the logic of designing, conducting and reporting school inspections is quite different from the logic of designing, conducting and reporting of research. When the school inspection logic is applied to answer research questions, this must lead to failure.

To trust or not to trust?

Assessment is an intrinsic part of the pedagogical process of learning and teaching, with two essential, complementary functions, namely to enhance and support learning, and to measure, as fairly as possible, the outcomes of learning. Although the two functions differ in certain respects, it needs to be emphasized that they rest on the same basic principles of validity, reliability, transparency and respect. Marking tests, i.e., evaluating and assigning points or grades to different qualitative levels of performance, is a crucial aspect of this, requiring a high degree of assessment literacy and credibility, or trustworthiness, among those who do the job.

The aim of the current paper is to highlight the issue of trust, in this case exemplified by the question whether ratings, and thereby also raters, can be trusted. However, our analysis of the SSI studies show that the issue of trust also must concern the methods and procedures used to investigate the trustworthiness of the raters. Because the SSI has relied on the logic of school inspections rather than on the logic of research to investigate trustworthiness of raters, the results they present are uninformative at best and untrustworthy at worst. The results also are interpreted within the logic of school inspections as showing the existence of severe deviations from the rules and regulations by certain schools, which results are loudly and widely disseminated. This creates distrust, without identifying ways to remedy the shortcomings,

This in turn carries the risk that simplistic, unfounded, or illogical solutions are resorted to, which needs to be avoided at all costs. There is no evidence to show that the system of national assessments would become more reliable if the responsibility for marking were to be handed over to somebody else than the individual student's teacher, be it a colleague or a central marker; nor would the system benefit from narrowing the construct, i.e. reducing what is assessed to ensure increased inter-rater consistency. On the contrary, the latter would lead to what is generally referred to as construct under-representation, which is considered a major threat to validity (Messick, 1989).

Furthermore, it needs to be emphasized that public naming and shaming of schools, and thereby of teachers' competence to mark their students' performances in a professional way, does not seem to be an ethical way of handling the complex situation of assessment and grading. This holds true in general, but it is perhaps even more relevant in the Swedish school system, which relies heavily on teachers' judgments at different levels. True, the national testing system provides considerable support, even though it could be clarified what weight this support is meant to carry, and to what extent double marking should take place.

Finally, there is an even more serious aspect to the question 'To trust or not to trust?', namely to what extent students can actually trust the assessment and grading made in schools, by teachers. This is crucial, not 'only' for personal and

pedagogical reasons, but also from an ethical, and juridical, point of view. To ensure the best outcome possible, all efforts need to be focused on enhancing the quality of assessment, not by shallow analyses, sweeping generalizations and quick fixes, but by serious, collaborative work based on solid research and reliable experience.

References

Erickson, G. (2009). *Nationella prov i engelska – en studie av bedömarsamstämmighet.* [National tests of English – a study of inter-rater consistency]. Retrieved 26 September 2011 from http://www.nafs.gu.se/publikationer

Erickson, G. & Åberg-Bengtsson, L. (forthcoming). A Collaborative Approach to National Test Development. In D. Tsagari & I. Csepes, *Collaboration in Language Testing and Assessment.* Frankfurt: Peter Lang Verlag.

Gustafsson, J.-E., & Myrberg, E. (2011). School inspections of Swedish schools: A critical reflection on intended effects, causal mechanisms and methods. Manuscript.

Harlen, W. (2004). Can assessment by teachers be a dependable option for summative purposes? In *Perspectives on Pupil Assessment – A paper presented to the GTC conference New relationships: Teaching, Learning and Accountability.* London, 29 November 2004. Retrieved 26 September 2011 from http://www.gtce.org.uk/policy/?year=2004&view=Filter

Jayasinghe, U. W., Marsh, H. W., & Bond, N. (2001). Peer Review in the Funding of Research in Higher Education: The Australian Experience. *Educational Evaluation and Policy Analysis*, *23*(4), 343-364.

McKinstry, B. H., Cameron, H. S., Elton, R.A. & Riley, S. C. (2004). Leniency and halo effects in marking undergraduate short research projects. *BMS Medical Education 4*(28).

Messick, S. A. (1989). Validity. I R. L. Linn (ed.), *Educational Measurement* (Third edition, pp. 13-103). New York: American Council on Education/Macmillan.

Skolinspektionen (2010). *Kontrollrättning av nationella prov i grundskolan och gymnasieskolan.* [Control marking of national tests for comprehensive school and upper secondary education]. Retrieved 26 September 2011 from http://www.skolinspektionen.se > Publikationer.

Skolinspektionen (2011). *Lika eller olika? Omrättning av nationella prov i grundskolan och gymnasieskolan.* [Remarking of national tests for comprehensive school and upper secondary education]. Retrieved 26 September 2011 from http://www.skolinspektionen.se > Publikationer.

Skolverket (2008). *Central rättning av nationella prov.* [Central marking of national tests]. Stockholm: Skolverket.

Skolverket (2009). *Bedömaröverensstämmelse vid bedömning av nationella prov* [Inter-rater consistency in marking national tests]. Stockholm: Skolverket.

SOU (2007:11). *Tydlig och öppen. Förslag till stärkt skolinspektion.* [Transparent and open. Reinforcing school inspection, commission report]. Stockholm: Utbildningsdepartementet.