Taking test results "into consideration" when grading

Anders Jönsson & Alli Klapp
AERA 2021

In Sweden, where this study is situated, grades are high stakes for students. Grades are the only criteria used for selecting students as they leave compulsory school and apply for upper-secondary school. When students apply for higher education, selection is also made based on the "Swedish Scholastic Aptitude Test", but a minimum of one third of the seats (often more) are based on grades. Given that it is the individual teacher who synthesizes students' performances into a grade, and that the reliability of teachers' grades has been questioned (e.g., Swedish National Agency of Education (SNAE), 2019), this is a problematic situation which can potentially have a significant influence on the lives of thousands of students each year.

Measures have been taken to increase the agreement in teachers' grading, most recently by legally requiring that teachers in primary and secondary education take results from national tests "into consideration" when grading. This relatively loose strategy of taking test results "into consideration" has not yet yielded any observable change in teachers' grades, however, and there is still a large discrepancy between teacher-assigned grades and test results for most subjects (SNAE, 2019)[1]. For example, the proportion of students whose grades correspond to their test results varies between 61-73 percent in the various subjects. Furthermore, it is more common that the teacher-assigned grades are higher than the test results, as compared to the other way around (with the exception of English as a foreign language, EFL). There are also gender differences, where girls to a slightly higher degree than boys receive grades that are higher in comparison with their test results. There are also large variations between schools.

The question in focus here is how this difference between teacher-assigned grades and test results originate and why it persists, which is investigated from a teacher perspective.

[1] No data is available for 2020, as the national tests were cancelled due to the Covid-19 pandemic.

Background

Grades and grading in Sweden

Swedish students are currently graded from year 6 and onwards. In the Swedish system, grades are standardized assessments regulated through the Swedish Education Act (2010:800), where it is stated, among other things, that grades must be decided by the teacher (or teachers) who conducted the teaching (Chapter 3, Section 16). The Education Act also states that the grades must be awarded on the basis of specific "knowledge requirements" (Chapter 10, Section 20), which are performance standards for the grades E (lowest passing grade), C, and A (highest passing grade). There are no explicit requirements for the failing grade (F), or for the "in-between-grades" D and B. The latter grades are to be awarded when the students have fulfilled all of the requirements of the lower grade and the majority of the requirements for the higher[2]. The knowledge requirements are found in the national curriculum for Swedish compulsory school ("Lgr11", Chapter 5) (SNAE, 2018).

Swedish teachers' grading practices have been investigated in a number of different studies and Korp (2006) identified three different approaches to grading through interviews with teachers (cf. Sadler, 2005). One approach involves a quantitative aggregation of students' test scores ("arithmetic approach"), while another involves making an overall decision based on a mix of data on student performance and other factors, such as personality traits, effort, and behavior. The existence of this latter ("intuitive") approach has been verified in other Swedish studies as well, both small-scale qualitative (e.g., Selghed, 2004) and large-scale quantitative studies (e.g., Klapp-Lekholm, 2008). That teachers tend to combine data on student performance with other factors when grading, particularly effort, is also a well-established phenomenon in international research (e.g. Brookhart et al., 2016; Malouff & Thorsteinsson, 2016), contributing to the low agreement among teachers that has been extensively documented.

As noted by Korp (2006), teachers using the arithmetic or the intuitive approach to grading did not make any references to the national grading criteria when talking about their grading practices. In contrast, some teachers used a third approach and compared data on student performance with the grading criteria, which is how the Swedish grading system is designed to work. In a later publication, Jönsson and Balan (2018) make a further distinction within this approach, by describing, first, an approach where the teachers grade individual assignments, where these "assignment

[2] For example, a student would be awarded a D if she/he fulfilled all of the requirements for an E and the majority, but not all, of the requirements for a C.

grades" are then used in the grading process (called "analytic grading"), and, second, an approach where the teacher makes an overall assessment of all existing data on student performance, without having any graded assignments to start from (called "holistic grading"). These authors also show, through an experimental comparison between the analytical and holistic grading approaches, that grades based on previously graded assignments have a higher agreement among teachers, as compared to holistic grading. The agreement is relatively low in both cases, however, with 66 and 46 percent agreement between the teachers in each group, even though the teachers agreed on both which criteria to use and the internal ranking among the students. Furthermore, the teachers in this study did not have any relationships with the students being graded, nor were they exposed to any form of external pressure from school leaders or legal guardians. Since the teachers used the same student performance for their grading, the variation within each group (i.e., analytic and holistic) is likely to depend on how the teachers chose to weigh the criteria and/or assignments.

Different paradigms

Swedish legislation requires teachers to take test results "into consideration" when grading. However, while teachers grading practices have been described as "complex, intuitive and tacit" (Bloxham et al., 2016, p. 466), test results are produced under quite different circumstances. This means that the teachers have to integrate information produced in a different context (i.e., testing situations), and from different assumptions about what assessment is all about, when making their decision.

The complex, intuitive and tacit nature of grading clearly relates to the concept of "qualitative judgment", as introduced by Sadler (1989). According to Sadler, qualitative judgments (or, more recently, "evaluative judgments", Boud et al., 2018) are made through the use of criteria and typically multiple criteria are used simultaneously when appraising the quality of student performance. Such judgments must be made by knowledgeable persons and cannot be reduced to a formula applied by nonexperts. In this view, the role of the assessor is similar to a "connoisseur", who "critiques" student performance by making comprehensive descriptions of the qualities and shortcomings of the work in the capacity of her/his expertise (Eisner, 1991). Without going into details about test design, it could be noted that this "assessment-as-judgment" view of assessment differs in several important aspects from a psychometric understanding of assessment.

For example, in a psychometric understanding of assessment, assessment is an "inferential activity" (Cizek et al., 2019). This means that the goal of assessment is to draw conclusions about

something that is not visible to the naked eye (i.e., student knowledge, learning, ability, etc.). Since student knowledge is not visible, these conclusions have to rely on "indirect samples of information" (Cizek et al., 2019). Within this paradigm, each item on a test is therefore used as an indication of a latent (invisible) trait, and more items generally means that more accurate conclusions can be drawn from the test scores. However, the information provided by individual items must be synthesized in order to make sense. In contrast to this view, evaluative judgments focus on the quality of performance. By judging the quality of an essay, a report, or other kind of extended task, the assessment is "direct" (Frederiksen & Collins, 1989), which means that no inferences *have to be made* about the student's knowledge or other latent traits. Similarly, when assessing the quality of a bowl of soup or an essay, no inferences need to be made about the cook or author (although, of course, such inferences can be, and often are, made in practice).

An obvious advantage of the directness of evaluative judgments, is that the assessment can be communicated without "translation". Test results, on the other hand, need to be translated from outcomes in the shape of scores, percentages, or competency statements, which are based on the aggregated information from several items, to concrete descriptions of strengths and weaknesses in student performance. A drawback of evaluative judgments, however, is that the simultaneous use of multiple criteria is an obscure and tacit process, to some extent relying on idiosyncratic beliefs of individual assessors, even if guided by shared criteria. As a consequence, teachers may have problems articulating their decisions and different teachers also tend to disagree on which grade/mark to award, even when grading the same performance using the same criteria (e.g., Bloxham et al., 2016; Jönsson & Balan, 2018). This clearly differs from a psychometric perspective, where the aggregation of scores on individual items, or subtests, into a total score is based on agreed upon algorithms or rules.

Another important difference between a psychometric understanding of assessment and evaluative judgments, is that tests have the ambition to "measure" knowledge and/or learning, which means that the outcome is quantified, and the results generally expressed along some kind of scale. Test results from different individuals, or from the same individual at different occasions, can thus be compared. Evaluative judgments, on the other hand, are qualitative, which means identifying, for example, strengths and areas in need of improvement in student performance, which cannot be easily compared or summarized.

A final example of differences between a psychometric understanding of assessment and evaluative judgments, is that both the performance and the scoring of tests tend to be standardized to achieve comparability. By keeping the testing situation as neutral as possible, the results are thought to be generalizable to many different situations. Since the context is not of interest, the items in a test are often considered independent of each other. However, for evaluative judgments, the context may be fundamental, because no objective or context-free knowledge is thought to exist. In order to be able to assess students' performance, the assessment situations therefore need to be carried out in contextualized or authentic situations, where the desired qualities can be expressed. As an example, Jönsson (2020) compares evaluative judgments with the widely used metaphor by Robert Stake, where a guest tasting the soup is used as an analogy for summative assessment. Although different aspects of the soup (such as temperature, thickness, saltiness, and so on) can be distinguished, they need to be evaluated in relation to the whole, since qualities such as thickness and saltiness cannot be evaluated in isolation from the soup.

Taken together, in the light of the abovementioned differences between a psychometric understanding of assessment and evaluative judgments, the idea that the differences observed between test results and grades are, wholly or partly, due to teachers being reluctant to take the test results into consideration when grading, may be an over-simplification of the difficulties faced by the teachers.

Taking test results into consideration

Despite several evaluations of differences between test results and grades (e.g., the Swedish Schools Inspectorate (SSI), 2018; SNAE, 2019), which have examined, for example, differences between teachers' assessments and external assessors' assessments, and between municipal and independent schools, there is a lack of knowledge about how teachers proceed when considering results from national tests in relation to their grading. As an exception, Vallberg-Roth et al. (2016) explored how teachers (n = 18) handled the relationship between test results and grading in an (unpublished) interview study with teachers in different subjects (e.g., EFL and history), who taught either in year 9 or in upper-secondary school. In this study, several of the EFL teachers claimed to attach great weight to the results from the national tests when grading. For one teacher, the test results were "absolutely decisive" for students' grades, while another teacher relied "very heavily" on the test results. But there were also teachers, who described the test results more as "a hint".

Among the teachers in history, not everyone was familiar with the national tests, as the tests for upper-secondary school are not mandatory and some teachers had chosen not to use the voluntary tests. The teachers who taught in compulsory school, however, expressed that the national test was an important part of grading. For one of them, there was a clear difference between how the students performed on the test, as compared to other data on student performance, while other teachers claimed that the test results were usually consistent with other data. In the cases where there was a difference between test results and other data on student performance, everyone claimed to handle the situation in a similar way. If the test result was better, the teachers awarded the student a higher grade. However, if the test result was lower, the reasons for the poor test result were explored, for instance by investigating whether the student had run out of time during the test situation. If the discrepancy could not be explained on the basis of such an investigation, the test result was considered valid and the student's grade was revised in correspondence with the test result.

The teachers in EFL also emphasized that it was necessary to take into account whether the student's results on the test were trustworthy. Above all, this applied to students in need of some kind of supplementary support, where differences in the support provided during the testing situation, as compared to ordinary teaching situations, could mean that the student's results were not considered to the same extent when assigning grades.

As is obvious from the findings presented by Vallberg-Roth et al. (2016), the relative weight attached to the test results may differ greatly among teachers, even among teachers teaching the same subject and the same grade level. It is also obvious that the teachers sometimes disagree with the test results, and that they have strategies to handle such disagreements, for instance by awarding the grade corresponding to the test results if the student manages to perform better on the test as compared to other, previous, assessment situations.

Although not emphasized by the authors, there are also some other interesting remarks that can be made from the interviews with these teachers. First, the teachers seem to compare the test results and the grades that they have awarded themselves on a one-to-one basis. As a consequence, the test results are only considered "better or worse" than the teacher-assigned grades[3], but not different in any other sense, such as in terms of coverage or alignment with the national curriculum. The comparison therefore appears unidimensional and without many nuances.

[33] In Sweden, both the results on the national test and the grades are expressed on the same scale (i.e., A-F).

A second remark is that the teachers seem to evaluate whether the test results are valid (or "trustworthy") only in cases where there is a difference in relation to their own decision, and when test result are lower than expected. Furthermore, the decision on whether to trust the test results relies primarily on factors related to the student, such as the student running out of time or having a "bad day", while other factors related to, for instance, test design, are not considered. Similar to the point made above, where grades and test results are compared on a one-to-one basis, if the test results are found not trustworthy, the entire test may be discarded for that student, instead of analyzing it there are any particular items or subtests that may have contributed to the invalid results.

Conclusions from previous research

Taken together, previous research suggests that Swedish teachers' grading practices primarily align with the evaluative judgment paradigm (Jönsson, 2020), although some teachers may rely more heavily on implicit rather than explicit criteria (Klapp-Lekholm, 2008; Korp, 2006; Selghed, 2004). As shown by Jönsson and Balan (2018), teachers may use a shared set of criteria when assessing student performance, even if they use different words to denote the qualities assessed. If some of the factors that can interfere with teachers grading are removed, as in the experimental design in the study by Jönsson and Balan (2018), teachers seem to agree on which criteria to use, the quality of student performance (i.e., strengths and areas in need of improvement), and the internal ranking among the students. When synthesizing these criteria into a grade, however, the agreement among teachers is low in all conditions. The findings thus suggest that it is mainly in this second, summarizing, step that teachers' judgments diverge, although it cannot be excluded that other factors, such as relationships with students, may interfere with the assessment during "normal conditions".

Previous research also provides some indications about how teachers handle differences between test results and grades, for instance by raising students' grades if they perform better than expected on the test, or by analyzing the validity/trustworthiness of the test results if the students perform less well. An interesting observation is that the teachers in the study by Vallberg-Roth et al. (2016) tend to compare test results and grades on a one-to-one basis, thereby disregarding the different ways in which the data on student performance has been collected (e.g., a single, standardized "one-shot" testing situation vs. data from regular classroom activities collected over time) and the different assumptions guiding the assessment (i.e., psychometric principles vs. evaluative judgment).

What is not known from previous research, however, is why these differences occur in the first place, or which aspects of validity/trustworthiness that teachers consider when analyzing the test results. The purpose of this study is therefore to explore the reasons for discrepancies between results on national tests and grades, and how teachers handle these discrepancies.

Method

The study is based on semi-structured interviews with teachers, which have been transcribed and analyzed with thematic content analysis.

Sample

The request for participation in the study has been directed to teachers who teach either physics or EFL in year 9. Teachers teaching year 9 have been selected, as it is during this year that the final grades are awarded for the Swedish compulsory school, and consequently the year when teachers are expected to take the results from the national tests into consideration. Although there are national tests in year 6 as well, these grades are not high stakes, as they are not used for any selection purposes.

The subject of physics has been chosen because there is a relatively large proportion (8 %) of teachers in physics who score national tests significantly higher as compared to external assessors (SSI, 2018). Physics is also one of the subjects, according to an analysis made by the Swedish National Agency for Education (2019), in which factors such as gender, migratory background, and school category (municipal or independent) have been seen to impact on teachers' grading.

EFL has been chosen, partly because it is a subject with a different character than physics, which means that these subjects are usually taught by different teachers, and partly because there is a relatively large proportion of EFL teachers (11 %) who score national tests significantly higher as compared to external assessors (SSI, 2018). According to the Swedish National Agency for Education (2019), EFL is also an exception, in the sense that it is more common for students in this subject to receive a grade that is lower than the test results, as compared to other subjects.

To get in touch with teachers who wanted to participate in the study, a request was posted via social media (Twitter and Facebook groups for teachers), at the same time as a number of key persons, who work with professional development for teachers, were asked to distribute the request through their channels. Via these announcements, several notifications were received, and the requests were removed when 20 teachers had responded, as this was considered a reasonable number

of respondents in relation to the scope of the project. However, one additional teacher contacted the researchers after the request was removed, asking to participate, which means that the sample consists of a total of 21 teachers who teach in year 9, either in physics (n = 9) or EFL (n = 12). Of these, 18 are women (86 %) and 17 work in municipal schools (81%). The teachers have a professional experience of on average 20 years (9-40 years).

Teachers from both municipal and independent schools are represented in the sample, as the analysis by the Swedish National Agency for Education (2019) suggests that independent schools on average give higher grades in relation to the test results. However, given the limited sample, there have been no ambitions to compare the responses from teachers who work at municipal or independent schools.

The interviews

Interviews was judged to be a reasonable method for the study, given that the purpose focuses on how the teachers justify discrepancies between test results and grades, and how they handle these discrepancies. Furthermore, semi-structured interviews were chosen, followed by a thematic content analysis (see below), as there were no pre-formulated categories to start the analysis from. The interviews, which followed an interview guide, were conducted over the Internet and documented in the form of audio files, which on average were about 23 minutes long (14-41 min.). All audio files were transcribed verbatim and the written transcripts were used as data in the study.

Analysis

The interview transcripts were analyzed with thematic content analysis, which is a method for identifying, analyzing, and interpreting patterns (or "themes") in qualitative data (Clarke & Braun, 2017). The analysis followed the procedure described by Braun and Clarke (2006), which in this case means that the following steps have been taken:

1. All transcripts were read thoroughly, and statements related to how the teachers handled discrepancies between test results and data on student performance were coded.
2. The codes were grouped in preliminary themes and data relevant to each theme was collected.
3. Preliminary themes were checked against coded statements and the data material as a whole, after which final themes were formulated.

4. A selection of examples of statements was made to support each theme. The selection was also made to show a breadth of statements and to represent the teachers' different voices in the selection.

Ethical considerations

All teachers participating in the study have received information about the purpose of the study and they have participated voluntarily. The interviews have only dealt with the teachers' professional practice and their consent is documented as audio files. No personal information other than names has been collected and quotes from each teacher are provided only in the form of an anonymous code. The code key is password protected and stored together with the audio files from the interviews on an external hard drive, which is locked. All participants were given preliminary interpretations of the interviews and had the opportunity to comment on them. The data has only been used for this study and new consent will be required before it is used in other contexts.

Results

The results from the analysis are presented below in the form of two themes, one broad theme called "Test results are misleading", which includes a number of subcategories, and one smaller theme called "Deficiencies are corrected". Quotes from teachers are followed by a code, which indicates the subject (E = EFL, P = physics), type of school (I = independent, M = municipal), and a number that indicates which individual is referred to.

Test results are misleading

The reason why students' grades differ from the results on national test is, with very few exceptions, that the test results are considered misleading by the teachers. However, there are several different reasons why the test results are considered misleading. Some depend on how the tests are constructed, others on the instructions for the tests, or on the students. In cases where the design of the tests contributes to the test results being considered misleading, the perceptions differ between the teachers of English and physics.

The tests are compensatory, grading is not

With regard to the test in English, all teachers think that the principle for summarizing the scores from the subtests into a total score differs too much from the grading in general, and that the compensatory principle in this weighing procedure tends to hide the shortcomings of some students:

> It is very forgiving, I must say, when I summarize the test score in English. So, if you are very weak in… have had a very weak performance in one of the subtests, it can almost disappear in the total score. (EM5)

> But this particular scoring system, yes, but that you got the test result E on ... or… yes, but E on the national test as a whole, it does not mean that you actually passed all the subtests, and in reality maybe should have an F. So, the whole national test thing … I think it is a bit misleading from a student perspective. (EM9)

A consequence of the principle for summarizing the scores from the subtests into a total score is that some students may pass the tests, even if they fail one of the subtests. If the student has not reached the minimum requirements for a passing grade according to the teacher's own documentation either, the teacher may choose to fail the student, even if the student has passed the national test. This, however, can create difficulties in the communication with students and legal guardians. It also means that there is a discrepancy between the grade and the test results, which is visible in the statistics. As mentioned above, EFL is the only subject where a relatively large proportion of students receive a lower grade as compared to the test results (SNAE, 2019).

Sampling from the curriculum

Another reason why the students' grades in EFL differ from the test results, which is also raised by all the teachers in the sample in one way or another, is that the test does not cover the full breadth of the curriculum. Some teachers describe, for example, how they usually work process-oriented in their teaching, by letting the students process their texts over time, which they are not given the opportunity to do on the test:

> I work very process-oriented, so that we… just like in real life, that you not only write a text and submit it, but you have to process it. This is not possible on the national tests. Those

> students who have strategies for processing their texts, putting them away, thinking about them, revising them and things like that, they do not get that opportunity. There are a lot of students who usually fail specifically in writing. (EM10)

Another example is that students, according to the curriculum, are expected to be able to write different kinds of texts, in different genres, and to different recipients, which is difficult to cover on the tests. A common point of view is also that the tests do not cover how the students are able to reflect on "living conditions, traditions, social relations and cultural phenomena in various contexts and areas where English is used" (SNAE, 2018, p. 36), even though this is a central part of the national curriculum:

> There is a knowledge requirement that is not tested there [on the tests], which is, perhaps a bit casually, called "factual knowledge". This is not included. And the requirement of using different types of sources and so on is not included either. (EM5)

> Yes, but they do not cover this digital reading. And they do not cover… but this may be different now when it [the test] will become web-based, so it could be that they produce that kind of items. Then they also do not cover this ability to reflect on and compare living conditions. (EM6)

Alignment with the curriculum

The teachers in both subjects generally think that there is a proper alignment between the tests and the national curriculum, and that the test results adequately represent the students' knowledge in relation to the knowledge requirements. There are exceptions, however, and some physics teachers think the tests are quite awful. What is criticized is that the students' factual knowledge does not have enough influence on the test results. Instead, these teachers claim that some students succeed in "cracking the code" for what is rewarded according to the assessment guidelines, and thus get an inflated test result despite limited factual knowledge. Similar to the compensatory principle for the tests in EFL, this means that the student receives a higher test result than what is suggested by the teacher's other data on student performance. However, the opposite view is also present in the sample of teachers, where several physics teachers believe that an important reason why the test results are misleading for some students is that the tests focus too much on specific details, which

the students do not necessarily have fresh in mind. Consequently, there is a disagreement among the teachers, where some want factual and detailed knowledge to be rewarded to a greater extent on the tests, while others prefer that more general abilities are rewarded, since facts and details are easily forgotten. A majority, however, think that the results will be misleading if the students lose points because they do not remember isolated facts or details, since this is not what is emphasized in the national curriculum:

> So we have students who are… they are motivated to study, many students crack the code for delivering what is expected on the national tests, how to express themselves /…/ the requirements for the higher scores are so low that you do not need to know anything. (PI1)

> For the subtest focusing more clearly on factual knowledge, it [the result] depends more on whether they mention the word "atom" in their explanation or not, or they lose points because of that. But I feel that I could have asked a follow-up question to the student: "Yes, but how were you thinking?" (PM2)

Access to supplement support

When it comes to factors not relating to the design of the tests, the teachers in the two subjects address similar reasons why the test results are considered misleading. An important reason is that some students have access to supplemental support during ordinary teaching. If these students are not allowed to have the same support during the tests, the test results will be misleading. One of the EFL teachers, who works with at-risk students, strongly opposes to forcing students into test situations they are not able to handle:

> I work with these low-performing groups, with those who do not take a third language, but who take only Swedish and English. And they often have dyslexia, they often have other types of difficulties, and to sit in a test situation and read the huge amount of text, which it is for these students, without support. I think it's a bit like child abuse actually. Because otherwise they may have a text-to-speech application that helps them. /…/ Or if it is text on paper, they have me or a friend who can help them by reading the text aloud to them. And then they will sit on a national test without any support and read a huge amount of text that they may not be able to decode. (EM7)

The physics teachers also highlight the difficulties for newly arrived, immigrant students to pass the tests, since these students do not always have the vocabulary required. As an example, one teacher (PM5) mentions the word "jet ski", which was used in one of the tests. The lack of vocabulary can to some extent be compensated with access to a dictionary, but the possibilities for using dictionaries decrease as the amount of text increases.

Student-related factors

Another important reason why the test results are considered misleading, mentioned by the teachers in both subjects, is that the tests are done at specific occasions (and only then), which means that there are factors, including random factors, that may make the students perform differently at this particular occasion. One such reason is that some students become stressed:

> I mean, there are so many who feel stressed as soon as we come to an assessment situation. /… / They have a total blackout. (PM5)

> We cannot rely on a single test occasion. If the student gets there and has a really bad day… which you can have. That has to be taken into consideration. If I have a student who normally performs on an A-level on everything, but only gets a B on the national test, should this student not be able to get the grade A then? (EM11)

Other factors

There are also some reasons for discrepancies between test results and grades, which are only mentioned by individual teachers, such as that the tests are perceived to vary in difficulty from year to year, or that the test content may have leaked, which means that the teachers do not know if they can trust the result.

Failures are corrected

One reason why test results and grades may differ, which is not based on the results being considered misleading, is that the teachers work to correct shortcomings in the students' knowledge, which have been identified with the help of the tests. If this work is successful, the students receive a passing grade, even if they failed the test. The work of correcting shortcomings is made possible

by the tests being performed as early as the beginning of April, so that there is a couple of months left of the spring semester, which can be used for focused teaching:

> And fortunately, the tests in science are not done in May, like the tests in math, which is much harder to correct. (PM1)

> You continue your work after national tests. It's not finished. (PM5)

Discussion

The purpose of this study was to investigate the reasons behind discrepancies between the results on the national tests and teacher-assigned grades from a teacher perspective, and how teachers handle these discrepancies. Interviews have therefore been conducted with teachers in two subjects (EFL and physics), which in evaluations by the Swedish Schools Inspectorate (2018) have been shown to have a relatively large discrepancy between test results and grades.

The interviews show that almost all cases of deviations between test results and grades are due to the teachers considering the test results to be misleading. This means that the teachers deliberately neglect the test results, instead of taking them into consideration when grading. As can be seen from the teachers' statements in the interviews, the teachers make this choice because they do not fully accept some of the psychometric principles that guide the design of the tests. As an example, all EFL teachers in this sample oppose to the idea that a student may pass the tests as a whole, if this student has failed one of the subtests. From a psychometric perspective, this compensatory principle makes sense, as high-stakes decisions should not be made on the basis of individual subtests, with a potentially high degree of uncertainty (such as performance tasks). The concept of measurement error is, however, totally absent in the teachers' statements, as is the concept of sampling. In relation to the latter, the teachers oppose to the idea that the test results should have a too strong influence on students' grades, as important requirements from the national curriculum are not covered by the tests. Rather than being guided by these psychometric principles, the teachers follow the guidelines for grading, where students are required to fulfill *all* knowledge requirements (i.e., national performance standards) for a grade level in order to be awarded this grade. Grading, however, is usually based on a much larger and more heterogeneous set of data on student performance, which is collected over time. This data may also include performances that are not covered by the standardized tests, such as process-oriented writing or highly contextualized information on peoples' living conditions. The tests, on the other hand, are "one-shot" events,

where the standardized format restricts the kind of knowledge or skills that may be tested in a meaningful way. Still, the teachers seem to evaluate the merits of the tests based on the guidelines for grading.

Evaluating the merits of the tests based on the guidelines for grading makes sense when considering that Swedish teachers' grading practices primarily align with the evaluative judgment paradigm. In line with this paradigm, grading involves the simultaneous use of multiple criteria to synthesize a sometimes extensive amount of heterogeneous data on student performance into a grade. However, as suggested by Jönsson and Balan (2018), although this approach seems adequate for appreciating the quality of individual performances, there are considerable disagreement among teachers when assigning grades. As has been shown in other studies, both Swedish and international, the grades also tend to include a number of other factors, some of which are not related to student achievement (e.g., Brookhart et al., 2016; Klapp-Lekholm, 2008; Malouff & Thorsteinsson, 2016; Selghed, 2004).

What the findings from this study seem to suggest is that the teachers use a similar approach for both assessing student performance and for synthesizing students' collective work into a grade, but that this approach is inadequate for the latter. While the assessment of student performance can be direct, without making any inferences about student proficiency beyond the quality of the task at hand, grading involves the transformation of the qualities identified in several performances into a point on an ordinal scale. Furthermore, when assessing individual tasks, there is no need to consider questions about sampling. This becomes an issue, however, when summarizing student performance into a grade, since the grade should represent student performance in relation to the knowledge requirements. Questions of uncertainty also change focus when moving from the assessment of individual performances to grading. For example, when assessing the quality of performance, it is important that teachers apply the criteria in a consistent manner, but when synthesizing students' collective work, other aspects come into play, such as the weight of individual performances in relation to the whole.

The strategies that teachers may have in order to handle the different demands in assessments and grading are not clearly visible in this study, nor in previous research on teachers' grading practices, which could indicate that they are indeed tacit and intuitive (Bloxham et al., 2016). What is visible here, however, is that the teachers dismiss some of the psychometric strategies for handling issues of uncertainty and sampling, since they obviously do not appeal to the teachers, but without

(explicitly) considering corresponding strategies in line with the evaluative judgment paradigm. It is therefore not known whether, or how, the teachers address these issues when grading, which means that different teachers are likely to do this differently, as is reflected in the disagreement among teachers, or not at all.

Although the teachers dismiss some psychometric principles, and others are absent in their reasoning, certain concepts are present to some degree. For example, although not expressed in those terms, several objections from the teachers are based on what they consider as threats to the validity of the (interpretations and uses of the) test results. One example is what some teachers consider an over-reliance on specific facts and details, which means that the tests deviate from what they are announced to cover. Another example is that the tests may not adequately capture the knowledge and skills of certain groups of students, such as students with special-education needs, immigrant students, or students prone to become stressed in testing situations. There are also a few objections in relation to reliability, for instance that students' test results may be influenced by random events. Again, however, teachers' objections are mainly framed by the guidelines for grading, where a more heterogeneous set of data on student performance may be incorporated, without the need to consider implications for the interrater agreement in scoring, and where individual adaptations to the testing situation may be accommodated. Furthermore, there are no indications of the teachers considering their own influence as assessors on the assessment outcome. This means that when differences between test results and other data on student performance are identified, it is primarily the test performance that is being scrutinized, not the teachers' own assessments. It also means that when there is no difference between test results and other data on student performance, teachers tend to take this as a confirmation of their own assessments. Either way, teachers' own assessments are "protected" from scrutiny.

In line with the findings by Vallberg-Roth et al. (2016), another observation is that several teachers seem only to consider the test results holistically, rather than analyzing individual items or subtests. When comparing the test results with other data on student performance on a one-to-one basis, teachers tend to discard the entire test if there are signs of the results being misleading, instead of only discarding the items or subtests that are considered misleading. Such handling of the test results is probably promoted by the fact that both the test results and the grades are expressed on the same scale (i.e., A-F), which may facilitate such a simplified comparison.

Conclusions

In Sweden, measures have been taken to increase the agreement in teachers' grading, most recently by legally requiring that teachers in primary and secondary education take results from national tests into consideration when grading. This strategy has, however, not yielded any observable change in teachers' grades, and there is still a large discrepancy between teacher-assigned grades and test results for most subjects.

According to the teachers interviewed in this study, this discrepancy exists primarily because teachers sometimes consider the test results to be misleading. Under such circumstances, the teachers deliberately choose not to take the results into consideration when grading. Furthermore, according to the teachers, it is primarily the design of the tests that contributes to the results being considered misleading, but in different ways. The reasons for considering the test design contributing to misleading results, seem to be rooted in view on assessment-as-judgment (Jönsson, 2020), which appears to be adequate for individual assessments of student performance, but not for synthesizing student performance and transforming the qualitative judgment into a point on a scale. In order to do the latter, concepts relating to uncertainty and sampling need to be considered, which are not present in the teacher interviews. Even if such concepts are present in the grading practice of teachers, albeit tacit and intuitive, the extensive disagreement among teachers suggests that any such practices need to be made explicit in order to become a shared asset within the teacher community.

Implications for policy and practice

Requiring teachers to take test results into consideration when grading is not necessarily as easy as it seems. Since teachers' grading practices primarily belong to the assessment-as-judgment paradigm, the psychometric principles guiding the design of national tests may appear alien to them, and the teachers may therefore dismiss these principles and consider the test results as invalid. Furthermore, teachers do not mention any other strategies for handling questions about uncertainty and sampling during the interviews, which could guide their grading practices. This suggests that these strategies may be tacit and intuitive, or even absent. However, in order to develop a shared practice for how to transform individual assessments to grades, these strategies need to be articulated and negotiated among the teachers. Consequently, the main implication from this study is that strategies, which are in harmony with the assessment-as-judgment paradigm, need to be

developed by (or together with) teachers, so that grades – if they are to be high stakes for students – are reasonably consistent and fair.

Limitations and future research

This is an interview study with a limited number of informants, all of whom have long teaching experience. The informants have also volunteered, which means that there is a risk that this sample of teachers is not representative of teachers in general. The results from this study must therefore be interpreted in relation to this specific sample, where the purpose is to understand the reasons for differences between test results and grades that these teachers pay attention to, rather than giving a picture of how Swedish teachers generally handle these differences. In addition, some parts of the results differ between the teachers of EFL and physics, which calls for caution in generalizing the conclusions to other subjects.

An obvious recommendation for future research, which is in line with the findings and the implications outlined above, is to support teachers in identifying and explicating strategies for how to transform individual assessments to grades, which are in harmony with the assessment-as-judgment paradigm.

Funding and cooperation

This project has been carried out in collaboration with Kunskapsskolan i Sverige AB.

References

Bloxham, S., den-Outer, B., Hudson, J., & Price, M. (2016) Let's stop the pretence of consistent marking: exploring the multiple limitations of assessment criteria, *Assessment & Evaluation in Higher Education, 41*, 466-481.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*, 77-101.

Brimi, H. M. (2011). Reliability of grading high school work in English. *Practical Assessment, Research & Evaluation, 16*, 1-12.

Brookhart, S. M., Guskey, T. R., Bowers, A. J., McMillan, J. H., Smith, J. K., Smith, L. F., Stevens, M. T., & Welsh, M. E. (2016). A century of grading research: Meaning and value in the most common educational measure. *Review of Educational Research, 86*, 803-848.

Clarke, V., & Braun, V. (2017). Thematic analysis. *Journal of Positive Psychology, 12*, 297-298.

Jönsson, A. (2020). Definitions of Formative Assessment Need to Make a Distinction Between a Psychometric Understanding of Assessment and "Evaluative Judgment". *Frontiers in Education: Assessment, Testing and Applied Measurement, 5*(2). doi: 10.3389/feduc.2020.00002

Jönsson, A. & Balan, A. (2018). Analytic or holistic: A study of agreement between different grading models. *Practical Assessment, Research & Evaluation, 23*(12). Retrieved from

https://scholarworks.umass.edu/pare/vol23/iss1/12/

Klapp Lekholm, A. (2008). *Grades and grade assignment: effects of student and school characteristics.* Doctoral dissertation: University of Gothenburg, Sweden.

Korp, H. (2006). *Lika chanser i gymnasiet? En studie om betyg, nationella prov och social reproduktion.* [Equal opportunities in upper-secondary school? A study of grades, national tests, and social reproduction.] Doctoral dissertation: Malmö University, Sweden.

Kunnath, J. P. (2017). Teacher grading decisions: Influences, rationale, and practices. *American Secondary Education, 45*, 68-88.

Malouff, J. M., & Thorsteinsson, E. B. (2016). Bias in grading: A meta-analysis of experimental research findings. *Australian Journal of Education, 60*, 245-256.

Parkes, J. (2013). Reliability in classroom assessment. I J. H. McMillan (Red.), *SAGE Handbook of Research on Classroom Assessment,* sid. 107-123. Los Angeles, CA, London, New Dehli, Singapore, Washington DC: SAGE.

Sadler, D. R. (2005). Interpretations of criteria-based assessment and grading in higher education. *Assessment & Evaluation in Higher Education, 30*(2), 175-194.

Selghed, B. (2004). *Ännu icke godkänt. Lärares sätt att erfara betygssystemet och dess tillämpning i yrkesutövningen.* [Not yet passed. Teachers' ways of experiencing the grading system and its application in professional practice.] Doctoral dissertation: Malmö University, Sweden.

Swedish Schools Inspectorate (2018). *Ombedömning av nationella prov 2017 – Fortsatt stora skillnader.* [Re-scoring national tests 2017 – Still large differences.] Report 2017:342.

Swedish National Agency of Education (2018). *Curriculum for the compulsory school, preschool class and school-age educare.*

Swedish National Agency of Education (2017). *Skolverkets systemramverk för nationella prov*. [Swedish National Agency of Education framework for national tests.]

Swedish National Agency of Education (2019). *Analyser av likvärdig betygssättning mellan elevgrupper och skolor*. [Analyses of equal grading between groups of students and schools.] Report 475.

Starch, D., & Elliott, E. C. (1912). Reliability of grading high-school work in English. *The School Review, 20*, 442-457.

Vallberg Roth, A-C., Gunnemyr, P., Londos, M. & Lundahl, B. (2016). Lärares förtrogenhet med betygssättning. [Teachers' familiarity with grading.] Unpublished report. Malmö University, Sweden.