Analytic or holistic? A study about how to increase the agreement in teachers’ grading

Anders Jönsson, Andrea Balan and Eva Hartell

Faculty of Education, Kristianstad University, Kristianstad, Sweden; City of Helsingborg, Helsingborg, Sweden; Departement of Learning in STEM at School of Industrial Engineering and Management, KTH Royal Institute of Technology, Stockholm, Sweden

ABSTRACT

In Sweden, grades are used for selection to upper-secondary school and higher education, even though agreement in teachers’ grading is low and the selection therefore potentially unfair. Furthermore, measures taken to increase the agreement have not been successful. This study has explored how to increase agreement in teachers’ grading by comparing analytic and holistic grading. Teachers (n = 74) have been randomly assigned to two different conditions (i.e. analytic or holistic grading) in either English as a foreign language (EFL) or mathematics. Findings suggest that analytic grading is preferable to holistic grading in terms of agreement among teachers. The effect was stronger and statistically significant in EFL, but could be observed in mathematics as well. However, teachers in the holistic conditions provided more references to criteria in their justifications, whereas teachers in the analytic conditions to a much larger extent made references to grade levels without specifying criteria.

Introduction

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 synthesises students’ performances into a grade, and that the reliability of teachers’ grades has been questioned (e.g. Swedish National Agency of Education, 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 account when grading. This relatively loose strategy of ‘taking test results into account’ 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 (Swedish National Agency of Education, 2019). The current situation may potentially lead to political demands for tying grades even more closely to test results, which in turn may have other unwanted consequences. For example, research suggests that the most widespread consequence of high-stakes testing is a narrowing of the curriculum, where teachers pay less attention to (or even exclude) subject matter that is not tested (Au, 2007; Pedulla et al., 2003). Furthermore, in their review of research on the impact of high-stakes testing on student motivation, Harlen and Deakin Crick (2003) conclude that results from such tests have been found to have a ‘particularly strong and devastating impact’ (p. 196) on low-achieving students.

The question addressed by this study is therefore whether it is possible to increase the agreement in teachers’ grading by other means than national tests, such as specifying how teachers synthesise data on student performance into a grade.

**Background**

**Grades and grading**

The term ‘grading’, as used here, means making a decision about students’ overall performance according to (implicit or explicit) grading criteria. It follows from this definition that grades (as a product) are composite ‘measures’ (i.e. expressed along an ordinal scale) of student proficiency, based on a more or less heterogeneous collection of data on student performance. It also follows from the definition that grading, as a process, involves human judgement.

In Sweden, it is the responsibility of the individual teacher to synthesise students’ performances into a grade and the teacher’s decision cannot be appealed. However, in those subjects were there are national tests,¹ the results from these tests have to be taken into account when grading. Exactly how the test results should be taken into account, or their weight in relation to the student’s other performances, is not specified. Obviously, a great deal of trust has been vested in teachers’ capacity to grade consistently and fairly.

**Consistency in teachers’ grading**

As shown in the review by Brookhart et al. (2016), covering more than 100 years of research about assessment and grading, research on teachers’ grading has a long history. In this research, two findings are particularly consistent: (a) Although student achievement is the factor that above all others determines a student’s grade, grades commonly include other factors as well, most notably effort and behaviour, and (b) There is great variability among teachers regarding both the process and product of grading (Brookhart, 2013).

Regarding the inclusion of non-achievement factors when grading, this seems primarily to be an effect of teachers wanting the grading to be fair to the students, which means that teachers find it hard to give low grades to students who have invested a lot of effort (Brookhart, 2013; Brookhart et al., 2016). Including non-achievement factors is therefore a way for teachers to balance an ethical dilemma in cases where low grades are anticipated to have a negative influence on students.
The variation in scores, marks, and grades between different teachers, and also for individual teachers on different occasions, has been extensively investigated. Several of the recent reviews of research about the reliability of teachers’ assessment and grading make reference to the early studies by Starch and Elliott (e.g. Brookhart et al., 2016; Parkes, 2013), who compared teachers’ marking of student performance in English, mathematics, and history (Starch & Elliott, 1912, 1913a, 1913b). These authors used a 100-point scale, and teachers’ marks in English (n = 142), for example, covered approximately half of that scale (60–97 and 50–97 points respectively for the two tasks). In history, the variability was even greater than it was for English and mathematics. They therefore concluded that the variation is a result of the examiner and the grading process rather than the subject (for an overview, see Brookhart et al., 2016).

Interestingly, almost a hundred years later, Brimi (2011) used a design similar to that used in the Starch and Elliott study in English, but with a sample of teachers specifically trained in assessing writing. The results, however, were the same (50–93 points on a 100-point scale).

In his review on the reliability of classroom assessments, Parkes (2013) also turns his attention to the intra-rater reliability of teachers’ assessment. As an example, Eells (1930) compared the marking of 61 teachers in history and geography at two occasions, 11 weeks apart. The share of teachers making the same assessment on both occasions varied from 16 to 90 percent for the different assignments. The 90 percent agreement was an extreme outlier, however, and the others were clustered around 25 percent. None of the teachers made the same assessment for all assignments, and the estimated reliability ranged from .25 to .51. The author concludes that ‘It is unnecessary to state that reliability coefficients as low as these are little better than sheer guesses’ (p. 52).

A number of objections can be made in relation to the conclusions above, due to limitations of the studies. For example, as pointed out by Brookhart (2013), the tasks used by Starch and Elliott would not be considered high-quality items according to current standards. Rather, they would be anticipated to be difficult to assess and likely to lead to large variations in marking. Another limitation is that most studies are ‘one-shot’ assessments, where teachers are asked to assess or grade performances from unknown or fictitious students. While it may be argued that such assessments are more objective, they miss the idea that teachers’ assessments can become more accurate over time as evidence of student proficiency accumulates. Further, their assessments can potentially become more valid as teachers come to understand what their students mean, even if expressed poorly. Lastly, teachers do not always have access to assessment criteria, which also means that their assessments could be anticipated to vary greatly. Still, teachers’ assessments are not always sufficiently reliable even with very detailed scoring protocols, such as rubrics. In a review of research on the use of rubrics, Jonsson and Svingby (2007) report that most assessments were below the threshold for acceptable reliability. Brookhart and Chen (2015), in a more recent review, claim that the use of rubrics can yield reliable results, but then criteria and performance-level descriptions need to be clear and focused, and raters need to be adequately trained. Taken together, even taking into account the limitations of some individual studies, the majority of studies on this topic point in the same direction with regard to the variability of teachers’ assessments and grading.
Different models for grading

The underlying reasons for the observed variability in teachers’ assessments and grading are complex and involve both idiosyncratic beliefs and external pressure, as well as a number of other factors, such as subject taught, school level, and student characteristics (e.g. Duncan & Noonan, 2007; Isnawati & Saukah, 2017; Kunnath, 2017; Malouff & Thorsteinsson, 2016; McMillan, 2003; Randall & Engelhard, 2008). Due to this complexity, teachers’ grading is sometimes portrayed in almost mystical terms, such as when Bloxham et al. (2016) write that assessment decisions, at least at the higher education level, are so ‘complex, intuitive and tacit’ (p. 466) that any attempts to achieve consistency in grading are likely to be more or less futile.

One possible reason for the observed variability in teachers’ grading, which is the focus of this study, is that teachers use different models (or strategies) for grading. Korp (2006) has, in a Swedish context, described three different models, which will be called holistic, arithmetic, and intuitive.

In the arithmetic model, the grade is calculated as a sum or a mean based primarily on test results. The arithmetic model therefore requires that the teachers document student performance quantitatively (cf. Sadler, 2005). According to Korp, the teachers who used this model did not mention either the national curriculum or the grading criteria when talking about their grading practice.

In the intuitive model, students’ grades are influenced by a mixture of factors, such as test results, attendance, attitudes, and lesson activity. From these factors, the teacher may have a general impression of the student’s proficiency in the subject, and that, rather than the specific performance in relation to the grading criteria, will determine the grade.

In contrast to the arithmetic and intuitive models, in the holistic model the teacher compares all available evidence about student proficiency to the grading criteria and makes a decision based on this holistic evaluation. Since it is only in the holistic model that the teachers base their decision on explicit grading criteria, this is the only model in line with the intentions of the Swedish grading system.

The fact that the holistic model works in line with the intentions of the grading system does not mean that this model is easier for teachers to apply. On the contrary, the teachers in the study by Korp (2006) expressed dissatisfaction with the idea that they were expected to integrate different aspects of student performance. Not surprisingly, it is considerably easier to arrive at a composite measure of student proficiency when using a homogeneous set of data, such as points from written tests, as compared to the heterogeneous material in a portfolio (Nijveldt, 2007). This tension between a unidimensional or multidimensional basis for grading is therefore yet another instance of the reliability versus validity trade-off. While unidimensional data may result in more coherent and reliable grading, such data only represents a fraction of student proficiency in a subject. Multidimensional data, on the other hand, may provide a fuller and more valid picture of student proficiency but is more difficult to interpret and evaluate in a reliable manner.

In addition to the models presented by Korp (2006), Jönsson and Balan (2018) introduced yet another model, which is called ‘analytic’ and can be described as a combination of the holistic and arithmetic models. In this model, teachers compare student performance on individual tasks with the grading criteria, producing what is
referred to as ‘assignment grades’, which are then used – more or less arithmetically – when assigning an overall grade. The analytic model thereby resembles the holistic model by making reference to explicit grading criteria (i.e. not only test scores or a general impression), but reduces the complexity of the final synthesis by already having transformed the heterogeneous data into a common scale.

As suggested by Jönsson and Balan (2018), the reduction of complexity in the analytic model may contribute to increasing the consistency in teachers’ grading while preserving the connection to shared criteria. A disadvantage, however, is that each individual decision is based on a much smaller dataset, as compared to a holistic judgement, which takes all available evidence about student proficiency into account. Consequently, there is a risk that the analytic model may influence the validity of the grades negatively in terms of construct underrepresentation.

**Analytic versus holistic assessment and grading**

Analytic assessment involves assessing different aspects of student performance, such as mechanics, grammar, style, organisation, and voice in student writing. Alternatively, holistic assessment means making an overall assessment, considering all criteria simultaneously.

The most apparent advantage of using analytic assessments is that they provide a nuanced and detailed image of student performance by taking different aspects into consideration. By judging the quality of, for instance, different dimensions of student writing, both strengths and areas in need of improvement may be identified, which, in turn, facilitate formative assessment and feedback.

In contrast, holistic assessments focus on the performance as a whole, which also has advantages. For example, holistic assessments are likely to be less time-consuming and there are indications that teachers prefer them (e.g. Bloxham et al., 2011). Furthermore, focusing on the whole prevents teachers from giving too much weight to individual parts. This is an important argument, since there is a widespread fear that easy-to-assess surface features get more attention in analytic assessment than more complex features that are more difficult to assess (Panadero & Jönsson, 2020). Sadler (2009), for instance, argues that teachers’ holistic and analytic assessments rarely coincide, and implies that analytic assessments may therefore not be valid. Focusing on different aspects of student performance can therefore be seen as both a strength and a weakness with analytic assessments. This is because even if surface features do not receive more attention than other features, the analytic assessments still need to be partitioned into separate criteria. Consequently, there is a risk of assessments becoming fragmented and missing the entirety.

However, in the same way as analytic assessments risk missing the whole when focusing on the parts, holistic assessments risk missing the parts when focusing on the whole. There is also, similar to the intuitive model of grading, the risk of unintentionally including other factors than student performance or attaching weight to other criteria than assumed (Bouwer et al., 2017). For example, Tomas et al. (2019) used analytic rubrics as supplementary to holistic assessments in order to elicit weightings of the criteria used. Their analyses indicated the existence of a ranking of importance and contribution of different criteria to the overall marks,
which differed from the expectations and assumptions of the assessors. While the assessors were assumed to give priority to complex skills such as critical thinking, descriptions and explanations of concepts contributed more highly towards the overall mark in their assessments. Since analytic assessments specify the criteria to consider, this situation is probably more common for holistic assessments. The authors therefore argue that analytic and holistic assessments should not be seen as opposites, but rather as complementary.

Still, there are researchers who are quite sceptical of analytic assessments (e.g. Bloxham et al., 2011; Sadler, 2009). In their defence, it should be noted that these sceptics often work within systems that require teachers to numerically score student performance according to analytic criteria, and where the individual scores are arithmetically aggregated into a composite score, mark, or grade. Under such circumstances, where qualitative and quantitative aspects of assessment are mixed, the sum of the parts may very well depart from the teacher’s holistic judgement, depending on how the scoring logic or algorithm is designed, which may lead to frustration and dissatisfaction. It could be argued, however, that such specific applications should be kept apart from the basic principles of analytic and holistic assessments.

**Materials and methods**

This study started from the observation that although grades are high stakes for students in Sweden, agreement in teachers’ grading is low, potentially making the selection to upper-secondary school and higher education unfair. And even though measures have been taken to increase the agreement in grading, these measures have not been very efficient. Furthermore, tying teachers’ grading even closer to results from national tests may have other negative consequences. This study therefore aims to explore other routes for increased agreement in teachers’ grading by comparing the advantages and disadvantages of analytic and holistic grading models in terms of a) agreement between teachers and b) how teachers justify their decisions.

The overall design of this study is experimental, where a number of teachers have been randomly assigned to two different conditions (i.e. analytic or holistic grading) in either English as a foreign language (EFL) or mathematics (Table 1). The study builds on a previous pilot study (Jönsson & Balan, 2018) and uses a similar methodology.

Information about the project was distributed in two different regions of the country, where teachers volunteered to participate in the study. The teachers chose whether to participate in EFL or mathematics and were then randomly assigned to one of the two conditions. The study was performed entirely online and no personal data has been collected, only grades and written justifications from the teachers.

**Table 1. Participants and conditions.**

|          | Analytic grading | Holistic grading |
|----------|------------------|------------------|
| EFL      | 24               | 18               |
| Mathematics | 17               | 15               |
Procedure

In the analytic condition, the teachers received written responses to the same assignment from four students on four occasions during one semester (i.e. a total of 16 responses). The assignments all addressed writing in EFL or mathematics, but otherwise had different foci (Table 2). All responses were from students aged 12, but with different levels of proficiency in either subject. The responses were authentic student responses that had been anonymised.

The teachers were asked to grade the student responses within a week after receiving them and report their decisions through an internet-based form. At the end of the semester, teachers were asked to provide an overall grade for each of the four students, accompanied by a justification. The final grades and written justifications were used as data in the study.

In the holistic condition, participants were given all of the material on a single occasion so that they were not influenced by any prior assessments of the students’ responses. Similar to the analytic condition, they were asked to provide a grade for each of the four students and a justification for each grade.

Agreement between and among teachers

A common method for estimating the agreement between different assessors (i.e. inter-rater agreement) is by using correlation analysis. Depending on whether scores (a continuous variable) or grades (a discrete variable) are given, either Pearson’s or Spearman’s correlation may be used. Spearman’s correlation (\(\rho\)) is a nonparametric measure of rank correlation, which is suitable for ordinal scales, such as grades. Naturally, if letter grades are used, they need to be converted to numbers in order to perform a correlation analysis. In this study, the grade A (i.e. the highest grade) has been converted to 6 and F (fail) to 1. Since only rank correlation has been used, no assumptions regarding equal distance between numbers are needed.

A disadvantage of using correlation analysis is that the assessments of two assessors may be highly correlated even if they do not agree on the exact grade, only the internal ranking. Therefore, in this study Spearman’s correlation has been combined with measures of absolute agreement. Pair-wise comparison is a common measure of agreement (Jönsson & Balan, 2018), but it has some notable disadvantages. First, it does not take the possibility of the agreement occurring by chance into account. Second, all agreements are given equal weight, which means that a couple of assessors who agree, but whose

| Task no. | EFL | Mathematics |
|----------|-----|-------------|
| 1        | Writing a biography of a relative or a friend of the family. | Eight simple tasks in arithmetic. |
| 2        | Writing an argumentative text about food waste. | Four short problem-solving tasks. |
| 3        | Writing a text about what a friend should be like. | One open-ended task that could be solved in many different ways. (written response) |
| 4        | Writing a short text message (SMS) to someone you care (or cared) about. | One open-ended task that could be solved in many different ways. (oral response) |
assessments deviate strongly from the consensus of other assessors, still contribute to the overall agreement. Third, since the observed agreements are divided by the number of all possible comparisons between assessors, the estimation tends to be low and it could be argued to underestimate the agreement between teachers.

In this study, we have therefore used Fleiss’ $\kappa$ and what we call ‘consensus agreement’ as measures of absolute agreement (for an in-depth discussion of different measures, see Stemler, 2004). Fleiss’ $\kappa$ provides an estimate similar to pair-wise agreement but takes the possibility of the agreement occurring by chance into account. We have also used ‘consensus agreement’, which is based on the idea that teachers as a group, or a ‘community of practice’, should ideally share a common interpretation of the grading criteria and standards, which the individual teacher either agrees or disagrees with. By using the most frequent grade for each student (i.e. the typical value), agreement among teachers has been estimated by calculating the proportion of grades that agree with the central tendency. Although this estimate of agreement only takes into consideration the proportion of grades that agree with the majority, it is generally higher than estimates based on pair-wise comparisons.

Finally, the median absolute deviation (MAD) has been calculated as a measure of distribution, and the Mann-Whitney U test has been used to test for statistical significance.

**Justifications by teachers**

The justifications for the grades were subjected to both qualitative and quantitative content analysis. First, all words the teachers used to describe the quality (either positively or negatively) of students’ performance were identified. All words were coded as different nodes in the data, even if they referred to the same quality. This was done in order to recognise the full spectrum of teachers’ language describing quality.

Second, all nodes were grouped in relation to commonly used criteria for assessing either writing (i.e. mechanics, grammar, organisation, content, style, and voice) or mathematics (i.e. methods, concepts, problem solving, communication, and reasoning). In addition, some teachers in EFL made references to comprehensibility and whether students followed instructions and finished the task. Some (in both EFL and mathematics) also made inferences about students’ abilities or referred only to grade levels (i.e. not describing quality). Four additional criteria, called ‘Comprehensibility’, ‘Rigour’, ‘Student’, and ‘Only grades’, were therefore added to the categorisation framework. For an example of the categorisation procedure, see Figure 1.

In the quantitative phase, the frequency of teachers’ references to the different criteria was used to summarise the findings and make possible a comparison between the different conditions.

**Findings**

The estimations of agreement between and among teachers are summarised in Table 3. In EFL, the teachers in the analytic condition agreed on the exact same grade in two thirds of the cases, while teachers in the holistic condition agreed in three out of five cases. The kappa estimations suggest fair (holistic condition) to moderate (analytic condition)
The student has understood all of the information and is able to express herself in a simple language [Style] in 3 of 4 tasks. The student’s texts have an audience [Voice]. There are examples of introduction, ending, greetings, questions to the reader [Organisation]. The message reaches the reader [Comprehensibility]. The language is simple with short sentences [Style]. Sometimes paragraphs are missing, punctuation marks, but the spelling is good [Mechanics]. The student has some weaknesses in sentence building, and grammar is sometimes wrong [Grammar], but the student’s messages and opinions are discernible [Voice] and can be understood by an English-speaking person [Comprehensibility]. (Justification for Student 1; Teacher 2 in analytic condition, EFL)

Figure 1. Typical example of justification and categorisation (in square brackets). Translated from Swedish by the authors.

Table 3. Comparison of agreement and correlations between the conditions.

|                | EFL  | Mathematics |
|----------------|------|-------------|
|                | Analytic | Holistic | Analytic | Holistic |
| Fleiss’ kappa  | .401 | .317 | .273 | .195 |
| Spearman’s rho (Std. deviation) | .929 | .952 | .893 | .919 |
|                | (.040) | (.057) | (.110) | (.102) |
| Consensus agreement | 66.7 | 59.7 | 57.4 | 51.7 |
| (MAD)          | (.375) | (1.00) | (.250) | (.500) |

Table 3. Comparison of agreement and correlations between the conditions.

\[ \text{Median absolute deviation (MAD)} \]

agreement. The mean rank correlation was high in both groups. The distribution estimate (MAD) was clearly greater in the holistic condition as compared to the analytic condition.

In mathematics, the agreement was also higher in the analytic condition as compared to the holistic condition, and the kappa estimations suggest slight (holistic condition) to fair (analytic condition) agreement. The mean rank correlation was high in both groups, and the distribution estimate was greater in the holistic condition as compared to the analytic condition.

Taken together, the agreement for the analytic condition is higher as compared to the holistic condition for both subjects, and for all estimates, although the difference is only statistically significant for EFL (p < .01).

Justifications

All in all, the teachers in EFL made 787 references to quality indicators in their justifications (i.e. on the average, 18.7 references per teacher), using 64 different terms for describing these qualities. As can be seen in Table 4, the teachers in the holistic condition made slightly more references in relation to most qualities. Similarly, in mathematics, the teachers made 358 references to quality indicators in their justifications (i.e. on the average, 11.2 references per teacher), using 29 different terms for describing these qualities (Table 5).
Table 4. Summary of teachers’ references in relation to the criteria for students and conditions (EFL).  

|                  | Student 1 | Student 2 | Student 3 | Student 4 | Analytic (mean) | Holistic (mean) | Sum  |
|------------------|-----------|-----------|-----------|-----------|-----------------|-----------------|------|
| Mechanics        | 13        | 16        | 7         | 11        | 19 (0.8)        | 28 (1.6)        | 47   |
| Grammar          | 25        | 26        | 23        | 27        | 56 (2.3)        | 45 (2.5)        | 101  |
| Organisation     | 24        | 25        | 16        | 31        | 45 (1.9)        | 51 (2.8)        | 96   |
| Content          | 5         | 18        | 3         | 8         | 14 (0.6)        | 20 (1.1)        | 34   |
| Style            | 66        | 96        | 67        | 78        | 149 (6.2)       | 158 (8.8)       | 307  |
| Voice            | 6         | 20        | 5         | 7         | 22 (0.9)        | 16 (0.9)        | 38   |
| Comprehensibility| 30        | 5         | 28        | 12        | 39 (1.6)        | 36 (2.0)        | 75   |
| Rigour           | 4         | 10        | 13        | 17        | 17 (0.7)        | 20 (1.1)        | 37   |
| Student          | 10        | 4         | 10        | 10        | 15 (0.6)        | 19 (1.1)        | 34   |
| Only grades      | 4         | 3         | 3         | 8         | 16 (0.7)        | 2 (0.1)         | 18   |
| **Total (mean)** | **187**   | **223**   | **172**   | **205**   | **392 (16.3)**  | **395 (21.9)**  | **787** |

Table 5. Summary of teachers’ references in relation to the criteria for students and conditions (mathematics).  

|                  | Student 1 | Student 2 | Student 3 | Student 4 | Analytic (mean) | Holistic (mean) | Sum  |
|------------------|-----------|-----------|-----------|-----------|-----------------|-----------------|------|
| Methods          | 17        | 15        | 16        | 17        | 37 (2.2)        | 28 (1.9)        | 65   |
| Concepts         | 24        | 23        | 23        | 20        | 36 (2.1)        | 54 (3.6)        | 90   |
| Problem solving  | 8         | 5         | 7         | 7         | 11 (0.6)        | 16 (1.1)        | 27   |
| Communication    | 11        | 10        | 14        | 18        | 21 (1.2)        | 32 (2.1)        | 53   |
| Reasoning        | 6         | 5         | 10        | 7         | 12 (0.7)        | 16 (1.1)        | 28   |
| Student          | 8         | 9         | 9         | 6         | 16 (0.9)        | 16 (1.1)        | 32   |
| Only grades      | 18        | 14        | 16        | 15        | 48 (2.8)        | 15 (1.0)        | 63   |
| **Total**        | **92**    | **81**    | **95**    | **90**    | **181 (10.6)**  | **177 (11.8)**  | **358** |

Table 4 summarises teachers’ references in relation to the criteria in EFL. Overall, most references were made to style dimensions (appr. 40%). This is also the most nuanced category, with 20 different terms used to describe these dimensions. In comparison, there were 9 terms used in relation to content, which comes second. Table 5 provides the corresponding information for mathematics, where the use of concepts (e.g. ‘fractions’ or ‘circumference’) is the largest category (appr. 25%).

There are some differences between the conditions, where teachers in the holistic groups provided comparatively more references in relation to most criteria for both subjects. In EFL, the largest relative differences are in relation to mechanics and content, while in mathematics the largest relative differences are in relation to concepts and communication. The most striking difference, however, is that it is primarily teachers in the analytic condition who make references only to grade levels. Table 6 shows an

Table 6. Comparison of teachers’ references in relation to the criteria for student 4 (mathematics).  

|                  | Analytic | Holistic |
|------------------|----------|----------|
| Methods          | 10       | 11       |
| Concepts         | 7        | 13       |
| Problem solving  | 4        | 3        |
| Communication    | 9        | 9        |
| Reasoning        | 4        | 3        |
| Student          | 2        | 1        |
| Only grades      | 10       | 5        |
example where the groups make similar assessments of the students’ merits according to the criteria, but the analytic group provides significantly more references only to grades.

Discussion

This study aimed to compare analytic and holistic grading conditions by investigating the extent to which teachers agree on students’ grades, as well as how they justify their decisions, when using an analytic or holistic model of grading.

Comparing the two conditions

As suggested by previous research (Jönsson & Balan, 2018), the analytic grading model may reduce the complexity of grading, thereby increasing the agreement between teachers. The findings lend some support to this interpretation as there is a difference between the two conditions in relation to both measures of agreement (i.e. Fleiss’ kappa and consensus agreement), where the agreement is higher in the analytic conditions, although the difference is only statistically significant in EFL. The MAD values also show that the teachers in the holistic groups tend to deviate more from the consensus. The correlations, however, are very similar across the groups, suggesting that the teachers’ perceptions of the relative rank of the students are largely unaffected by the model of grading. This finding also suggests that there is indeed a shared conception of quality performance, although teachers may disagree on the exact grade.

In relation to teachers’ justifications, critics of analytic assessments tend to assume that such assessments result in construct underrepresentation (e.g. Sadler, 2009). This is to some extent verified by the findings, since teachers in the holistic conditions provided slightly more references in relation to most criteria and the teachers in the analytic conditions provided substantially more references to grade levels, without describing any qualities. Interestingly, in EFL, teachers in the holistic condition provided more references to mechanics, which is often seen as a surface feature, but also to content. Neither of the conditions could therefore be accused of focusing primarily on surface features, such as organisation, mechanics, and grammar in EFL.

Regarding the assessment of individual students, there were no clear differences between the groups. Instead, there is quite a strong consensus about the qualities in students’ performances. For instance, as can be seen in Table 6, both groups agree that the use of methods is a strength for this student, as are communication and concepts, while reasoning and problem solving are relative weaknesses. This consensus means that there is a potential for the teachers to agree on the formative feedback to give the students (i.e. strengths and suggestions for improvement). Consequently, formative assessments may not necessarily be affected by the inequality of grading if no overall assessment has to be made.

In sum, the findings suggest that the teachers in the sample are in agreement about which criteria to use when assessing and also about the extent to which these criteria are fulfilled in students’ texts. The teachers also largely agree on the rank order of student performance. However, when assigning specific grades, the agreement is generally low. This observation supports the idea of assessment as a two-tier process, where the first stage involves the discernment of criteria in relation to the performance, and the second
involves making a judgement about the quality of the performance (Sadler, 1987). Teachers may therefore be in agreement during the first stage but not the second (or vice versa), for instance, because they attach different weight to individual criteria when making an overall assessment. The agreement was higher for the analytic model in both subjects and for both measures, but the difference was only statistically significant in EFL.

**Findings in relation to research on teachers’ grading**

It is difficult to draw any firm conclusions about teachers’ grading based on this study, since the conditions in the experimental situation differ in several respects from how teachers may work during ordinary conditions. There are, however, some interesting observations that can be made.

First, even if the agreement in the analytic conditions was higher as compared to the holistic conditions, agreement in all groups was relatively low (52–67%), which is in line with previous research on teachers’ grading (e.g. Brookhart et al., 2016). This could to some extent be expected, as the study uses a methodology similar to that used in several other studies, with tasks not designed by the teachers themselves and fictitious students. What might differ, however, is that the teachers in this study had access to shared grading criteria (which are very detailed in Sweden) and that the teachers in both subjects have experience with assessing student performance on national tests – both of which could be assumed to contribute to increased agreement. Furthermore, several of the factors that have previously been shown to influence teachers’ grading, such as student characteristics and external pressure (e.g. Malouff & Thorsteinsson, 2016; McMillan, 2003), were not part of the experimental conditions, which could also be assumed to lower the variability among teachers and increase the agreement. As can be seen, however, the agreement is still low, which might indicate that factors such as idiosyncratic beliefs (e.g. Kunnath, 2017) and individual weighting of criteria contribute more strongly to the variability.

Second, in the justifications provided by the teachers, there are some references to the students as persons, but otherwise there are no signs of teachers taking other aspects into consideration. This finding is most likely a methodological artefact, partly since, as discussed above, such aspects were not part of the experimental conditions and the teachers therefore had no personal information about the students, and partly because the teachers knew that someone was going to review their justifications. What this suggests, however, is that it may be possible for teachers to confine themselves to using only legitimate criteria if they know that their grading will be reviewed.

**Conclusions and implications for practice**

The findings from this study suggest that analytic grading, where teachers assign grades to individual assignments and then use these ‘assignment grades’ when deciding on the final grade, is preferable to holistic grading in terms of agreement among teachers. The effect was stronger and statistically significant in EFL, but a clear tendency could be observed in mathematics as well, for all measures.

Teachers in the holistic conditions provided more references to criteria in their justifications, whereas teachers in the analytic conditions to a much larger extent made
references to grade levels without specifying criteria. The gain in agreement for the analytic grading model is consequentially countered by teachers making fewer references to criteria. This effect is, however, quite small.

All groups were in agreement on which criteria to use when assessing student work and the qualities identified in students’ performance, which means that formative assessments may not be affected to the same extent by the low agreement in teachers’ grading. Furthermore, if adopting an analytic model of grading, teachers may consider keeping the ‘assignment grades’ to themselves, as these are based on limited data and could be assumed to be unreliable, while only communicating qualitative feedback to the students (i.e. strengths and suggestions for improvements) in order to support students’ learning and improved performance.

**Limitations and future research**

There are a number of important limitations in this study that need to be considered. First, although it is an experimental study where the participants were randomly assigned to the conditions, the sample was small and the participants volunteered to participate. Second, the experimental conditions differed in several respects from teachers’ ordinary grading; for instance, the teachers did not design the tasks themselves and had no personal relationship with the students. Furthermore, as mentioned above, the current study cannot confirm which individual and/or contextual factors influenced teachers’ grading, only that some factors beyond the grading model – such as giving different weight to different criteria – gave rise to variability in the sample. Given the variability of assigned grades, as well as the fact that this influence has been a robust finding in numerous studies over the years, the lack of support in this study is most likely an artefact of the design. Teachers can be assumed to restrict their written judgements to what they believe are legitimate criteria, since they know that someone will review their grading.

It should also be emphasised that no comparisons can be made between the two subjects since it is not known whether student performance is equally difficult to synthesise. For instance, only written performance was included for the teachers in EFL, while there was also oral performance for the teachers in mathematics, making the material more heterogeneous. Such differences may explain why the agreement between teachers in mathematics was lower as compared to the teachers in EFL in this study.

In conclusion, the aim of this study was to explore alternative ways to increase the agreement in teachers’ grades, as opposed to increasing the influence of test results. Findings suggest that the agreement between teachers may increase by adopting an analytic grading model, but only slightly. Furthermore, agreement between teachers is still low, even when a large number of factors, which have previously been shown to influence teachers’ grading, were removed as part of the experimental design. The results could therefore be used to support the argument that teachers’ grading is indeed so ‘complex, intuitive and tacit’ (Bloxham et al., 2016, p. 466) that any attempts to achieve consistency in grading are likely to be futile. On the other hand, the findings could be seen as a small step towards increasing the agreement. For example, what would happen if the teachers, in addition to adopting the analytic grading model, used task-specific
criteria for assessing the individual assignments, and then simpler criteria for aggregating the ‘assignment grades’, instead of using the (detailed, but abstract) national grading criteria? Or what would happen if the teachers, in addition to adopting the analytic grading model, participated in a moderation procedure, where they reviewed each other’s grading?

We propose a continued investigation into possibilities to increase agreement in teachers’ grading in ways that can be implemented as part of ordinary practice, such as the ones mentioned above (i.e. improving criteria and participating in moderation procedures), as well as other strategies suggested by previous research, such as aiding teachers in identifying different levels of quality (Brookhart et al., 2016). Such an endeavour can be supported by currently available resources (e.g. Brookhart & Nitko, 2008; Guskey & Brookhart, 2019) and may contribute to steering clear of a too-heavy reliance on external test results and the unfair grading that Swedish students experience today.

Note

1. There are currently tests provided in ten subjects during the last year of compulsory school, but each individual student only performs five of them.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Swedish Research Council under Grant [2018–03389].

Notes on contributors

Anders Jönsson is professor of Educational Sciences at Kristianstad University. His research interest is in classroom assessment, for both formative and summative purposes. Anders is the scientific leader of the research platform ‘Collaboration for Learning’, which aims to contribute to the development of teaching in schools and preschools through practice-based research.

Andreia Balan is an assistant professor at the City of Helsingborg. Her research interest is in formative assessment and mathematics education.

Eva Hartell is a teacher in STEM subjects and holds a PhD in the area of classroom assessment. She is currently working with research and development in the municipality of Haninge and at KTH Royal Institute of Technology in Sweden.

ORCID

Anders Jönsson http://orcid.org/0000-0002-3251-6082
References

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. Educational Researcher, 36(1), 258–267. https://doi.org/10.3102/0013189X07306523

Blokhams, S., Boyd, P., & Orr, S. (2011). Mark my words: The role of assessment criteria in UK higher education grading practices. Studies in Higher Education, 36(6), 655–670. https://doi.org/10.1080/03075070100377716

Blokhams, 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(3), 466–481. https://doi.org/10.1080/02602938.2015.1024607

Bouwer, R., Verhavert, S., Lesterhuis, M., van Gasse, R., Donche, V., & De Maeyer, S. (2017, November). Interpreting the validity of misfit statistics in comparative judgement. AEA Europe.

Brimi, H. M. (2011). Reliability of grading high school work in English. Practical Assessment, Research & Evaluation, 16(1), 1–12. https://doi.org/10.7275/j531-fz38

Brookhart, S. M. (2013). The use of teacher judgement for summative assessment in the USA. Assessment in Education: Principles, Policy & Practice, 20(1), 69–90. https://doi.org/10.1080/0969594X.2012.703170

Brookhart, S. M., & Chen, F. (2015). The quality and effectiveness of descriptive rubrics. Educational Reviews, 67(3), 343–368. https://doi.org/10.1080/0013189X.2014.929565

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(4), 803–848. https://doi.org/10.3102/0034654316672069

Brookhart, S. M., & Nitko, A. J. (2008). Assessment and grading in classrooms. Pearson.

Duncan, C. R., & Noonan, B. (2007). Factors affecting teachers’ grading and assessment practices. The Alberta Journal of Educational Research, 53(1), 1–21.

Eells, W. C. (1930). Reliability of repeated grading of essay type examinations. Journal of Educational Psychology, 2(1), 48–52. https://doi.org/10.1037/h0071103

Guskey, T. R., & Brookhart, S. M. (2019). What we know about grading: What works, what doesn’t, and what’s next. ASCD.

Harlen, W., & Deakin Crick, R. (2003). Testing and motivation for learning. Assessment in Education: Principles, Policy & Practice, 10(2), 169–207. https://doi.org/10.1080/0969594032000121270

Isinawati, I., & Saukah, A. (2017). Teachers’ grading decision making. TEFLIN Journal, 28(2), 155–169. https://doi.org/10.15639/teflinjournal.v28i2/155-169

Jonsson, A., & Balan, A. (2018). Analytic or holistic: A study of agreement between different grading models. Practical Assessment, Research & Evaluation, 23(12), 1–11. https://doi.org/10.7275/z3gm-fp34

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. Educational Research Review, 2(2), 130–144. https://doi.org/10.1016/j.edurev.2007.05.002

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

Kunnath, J. P. (2017). Teacher grading decisions: Influences, rationale, and practices. American Secondary Education, 45(3), 68–88.

Malouff, J. M., & Thorsteinsson, E. B. (2016). Bias in grading: A meta-analysis of experimental research findings. Australian Journal of Education, 60(3), 245–256. https://doi.org/10.1177/0004944116664618

McMillan, J. H. (2003). Understanding and improving teachers’ classroom assessment decision making: Implications for theory and practice. Educational Measurement: Issues and Practice, 22(4), 34–43. https://doi.org/10.1111/j.1745-3992.2003.tb00142.x
Nijveldt, M. J. (2007). Validity in teacher assessment: An exploration of the judgement processes of assessors [Doctoral dissertation]. Leiden University.

Panadero, E., & Jonsson, A. (2020). A critical review of the arguments against the use of rubrics. Educational Research Review, 30, 100329. https://doi.org/10.1016/j.edurev.2020.100329

Parkes, J. (2013). Reliability in classroom assessment. In J. H. McMillan (Ed.), SAGE handbook of research on classroom assessment (pp. 107–123). SAGE.

Pedulla, J. J., Abrams, L. M., Madaus, G. F., Russell, M. K., Ramos, M. A., & Miao, J. (2003). Perceived effects of state-mandated testing programs on teaching and learning: Findings from a national survey of teachers. National Board on Educational Testing and Public Policy.

Randall, J., & Engelhard, G. (2008). Differences between teachers’ grading practices in elementary and middle schools. The Journal of Educational Research, 102(3), 175–185. https://doi.org/10.3200/JOER.102.3.175-186

Sadler, D. R. (2005). Interpretations of criteria-based assessment and grading in higher education. Assessment & Evaluation in Higher Education, 30(2), 175–194. https://doi.org/10.1080/0260293042000264262

Sadler, D. R. (2009). Indeterminacy in the use of preset criteria for assessment and grading. Assessment & Evaluation in Higher Education, 34(2), 159–179. https://doi.org/10.1080/02602930801956059

Sadler, R. D. (1987). Specifying and promulgating achievement standards. Oxford Review of Education, 13(2), 191–209. https://doi.org/10.1080/0305498870130207

Starch, D., & Elliott, E. C. (1912). Reliability of grading high-school work in English. The School Review, 20(7), 442–457. https://doi.org/10.1086/435971

Starch, D., & Elliott, E. C. (1913a). Reliability of grading work in mathematics. The School Review, 21(4), 254–259. https://doi.org/10.1086/436086

Starch, D., & Elliott, E. C. (1913b). Reliability of grading work in history. The School Review, 21(10), 676–681. https://doi.org/10.1086/436185

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. Practical Assessment, Research & Evaluation, 9(4), 1–11. https://doi.org/10.7275/96jp-xz07

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 No. 475).

Tomas, C., Whitt, E., Lavelle-Hill, R., & Severn, K. (2019). Modelling holistic marks with analytic rubrics. Frontiers in Education: Assessment, Testing and Applied Measurement, 4(89). https://doi.org/10.3389/feduc.2019.00089