Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 23 Number 12, September 2018 ISSN 1531-7714

Analytic or Holistic: A study of Agreement Between Different Grading Models

Anders Jönsson, *Kristianstad University*
Andreia Balan, *City of Helsingborg*

Research on teachers' grading has shown that there is great variability among teachers regarding both the process and product of grading, resulting in low comparability and issues of inequality when using grades for selection purposes. Despite this situation, not much is known about the merits or disadvantages of different models for grading. In this study, a methodology for comparing two models of grading in terms of (a) agreement between assessors (reliability) and (b) justifications for the grades assigned (validity) was used with a small sample of teachers (n = 24). The design is experimental, with teachers being randomly assigned to two conditions, where they graded the same student performance using either an analytic or a holistic approach. Grades have been compared in terms of agreement and rank correlation, and justifications have been analyzed with content analysis. Findings suggest that the analytic condition yields substantively higher agreement among assessors as compared to the holistic condition (66 versus 46 percent agreement; Cohen's kappa .60 versus .41), as well as higher rank correlation (Spearman's rho .97 versus .94), without any major differences in how the grades were justified. On the contrary, there was a relatively strong consensus among most raters in the sample.

In most educational contexts, *grading* means making a holistic judgment about student overall performance according to grading criteria. Even though the criteria and the scale may differ, grading often involves taking a diverse set of performances – such as written tests, lab-reports, oral presentations, group discussions, etc. – into account when making a decision about students' final grades. Teachers may have very different strategies for this complex endeavor (e.g. Korp, 2006; McMillan, Myran, & Workman, 2002), which is manifested in the great variation in the grades teachers assign to students' work (Brookhart et al., 2016).

This study uses a combination of experimental design and content analysis to compare two different models of grading student performance: one analytic and one holistic. In the holistic model, teachers make a decision about students' grades from a holistic judgment of all available data on student proficiency in the subject (i.e. similar to a portfolio). They also refrain from making analytical judgments along the grading scale for individual assignments during the semester. In the analytic model, on the other hand, teachers continuously grade students' assignment and use these "assignment-grades" when deciding on an overall grade at the end of the semester. The main differences are therefore that in the analytic model, each decision is based on student performance on individual assignments and that the teachers use these "assignment-grades" to inform their decision on the final grade, while in the holistic model the decision is based on a more comprehensive set of data with no previous quantitative assessments to inform the decision. The purpose is to investigate whether there are differences between these models in terms of agreement among teachers and how teachers use data on student performance to inform their decisions.

Background

What are grades?

The term "grading", as used here, means making a holistic judgment about student overall performance according to 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 judgment.

Assessment (as a process) consequently differs from grading in that "assessment" refers to making a judgment about the quality of student performance on an individual assignment. Assessment as a product, however, may be expressed either along a scale or as a qualitative description of strengths and suggestions for improvements. In the latter case, the assessment can easily be used as formative feedback, while aggregated and codified information (such as scores or grades) is better suited for summative assessments, since such information may need to be transformed in order to serve as input in formative assessment.

Research on teachers' grading

Research on teachers' grading has a long history, not least shown by the review by Brookhart et al. (2016), covering over 100 years of research about assessment and grading. In this research, two findings are particularly consistent over the years: (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 behavior, 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). To include 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. That low grades can have a negative influence on subsequent performance is shown by, for instance, Klapp (2015). Klapp investigated how grading in primary school affected students' achievement in secondary school by comparing data from students who received grades in Grade 6 and students who did not (n = 8,558). The results showed a main significant negative effect of grading on subsequent achievement during secondary school. This effect was more pronounced for low-ability students, who also finished upper secondary school to a lesser extent, as compared to students not being graded during primary school.

The variation in scores, marks, and grades between different teachers, but also for the same teachers at 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 & Elliot, 1912; 1913a; 1913b). These authors used a 100 points 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, as compared to English and mathematics. They therefore conclude that the variation is a result of the examiner and the grading process, rather than the subject (for on overview, see Brookhart et al., 2016). Interestingly, almost a hundred years later Brimi (2011) used a similar design as the Starch and Elliott study in English, but a sample of teachers specifically trained in assessing writing. The results, however, were the same (50-93 points on a 100 points 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 at both occasions varied from 16-90 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 have been considered high-quality items according to current standards. Rather, they would have been anticipated to be difficult to assess and lead to large variations in marking. Another limitation is that most studies are "one-shot" assessment, where teachers are asked to assess or grade performances from unknown or fictitious students. While such assessments may be argued to be more objective, this procedure misses the idea of teachers' assessments becoming more accurate over time, as evidence of student proficiency accumulates, and potentially more valid since the teacher knows what her/his 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 (2014), 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 if acknowledging the limitations of individual studies, the amount of studies on this topic, where most point in the same direction, the variability of teachers' assessments and grading has to be considered a robust finding. Furthermore, this variability can be quite large. As an example, Kilday, Kinzie, Mashburn, and Whittaker (2012) report that 40 percent of the total variance in teachers' assessments could be attributed to differences between teachers.

The documented variability of teachers' assessments and grading raises the question where this variation comes from. This has turned out to be a complex and intriguing question and both quantitative and qualitative research have made efforts to understand teachers' grading practices. As an example of quantitative research designs, Duncan and Noonan (2007) showed, based on a survey of approximately 500 high-school teachers, that the subject taught influenced teachers' grading practices. Randall and Engelhard (2008), who measured teachers' responses to a number of scenarios describing different student characteristics, showed that the practices of elementary and middle-school teachers differed (where the former were generally more lenient). Obviously, there are a number of contextual factors that may influence teachers' grading. Furthermore, Malouff and Thorsteinsson (2016) present a meta-analysis of research findings on the existence of bias in the grading of student work, where a number of student characteristics are shown to result in lower grades. These characteristics are "students who have negative educational labels, students who are members of specific ethnic or racial groups, students who have previously performed poorly, and less attractive students" (p. 252).

As an example of qualitative research designs, Isnawati and Saukah (2017) performed in-depth interviews with two teachers from different junior high schools, showing that the teachers held strong beliefs that assigning grades was not only about accurately representing students' proficiency, but also for purposes of life-long learning and motivation. The finding that teachers' grading practice is influenced by idiosyncratic beliefs has been verified in a number of studies and helps in explaining the variability of teachers' grading practices. In particular, the research by James H. McMillan has contributed to the understanding of teachers' grading. In one of his publications (McMillan, 2003), he presents a model for teachers' decision making, which is seen as a process where teachers balance the demands of: (a) external factors (e.g. accountability and the influence of parents) and (b) constraints (e.g. the disruptive behavior of students) with their own beliefs and values to determine classroom assessment practices. This model has been used in subsequent studies, such as Kunnath (2017), who showed that teachers' grading were "strongly influenced by teachers' philosophy of teaching and learning, their concern for external perceptions, and administrator pressure on assigning low grades" (p. 85). Considering that both individual and contextual factors in the model may differ, it is no wonder that there is variation among teachers; a situation that has led researchers to suggest that grading practices may require more attention in teacher-education programs (e.g. Randall & Engelhard, 2010).

Different models of grading

Korp (2006) has, in a Swedish context, described how teachers use different models for grading, which will be called holistic, arithmetic, and intuitive[1].

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. This differs from the arithmetic model, in which the grade is calculated as a sum or a mean based on test results or grades on individual assignments. The arithmetic model therefore requires that the teachers document student performance as points or grades on tasks and tests. According to Korp, the teachers who used this model did not mention neither the national curriculum nor the grading criteria when talking about their grading practice.

The third model for grading is called the intuitive model and corresponds to the grading practice of teachers as discussed above (e.g. Gipps, Brown, McCallum, & McAlister, 1995; McMillan, 2003). In this 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, which will determine the grade, rather than the specific performance in relation to the grading criteria. For instance, Korp cites a language teacher with extensive experience, who believes that the grading criteria should become second nature to teachers and that she can "see" which students who will eventually receive higher grades.

Of these three models, it is only the holistic model that is in line with the intentions of the Swedish grading system, since the grading is done in relation to official criteria (and only in relation to these criteria). In the arithmetic model, on the other hand, teachers' grading has no clear connection to shared criteria. Furthermore, students' grades are based on a sum or mean, which means that students who have not met all the requirements can still pass, if their combined test scores exceed the cut-off score determined by the teacher. In the intuitive model, the relation to the grading criteria is also weak, since the grade is based on a general impression of the student. Furthermore, the grade is influenced by factors that are not included in the grading criteria. Nevertheless, this model is obviously widespread, both in Sweden and internationally.

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 Korp's (2006) study expressed a dissatisfaction with the expectations to integrate different aspects of student performance with each other. 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 instantiation 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.

A new model of grading

From the publication of Korp's (2006) study till today, the Swedish curriculum has undergone a major reform, among other things resulting in a new grading scale with 6 levels from A-F (as compared to 4 levels in the previous scale). This change has affected teachers' grading practices and also led to the emergence of a new model for grading, here called "analytic model", not identified by Korp. This model could be described as a hybrid between the arithmetic and the holistic models. It is arithmetic in the sense that teachers grade individual assignments according to the six-level grading scale, resulting in a number of "assignment-grades" (e.g. A, C, C, E, A). However, in contrast to the arithmetic model, these "assignment-grades" are assigned in relation to the grading criteria, similar to the holistic model.

The advantage of this new model is that it reduces the complexity of grading. By assigning grades to individual assignments, decisions are made based on less heterogeneous data. These "assignment-grades" are then used, more or less arithmetically, in order to inform the decision about the final grade. Hypothetically, this procedure could result in more reliable grading, while

[1] Korp's (2006) original categorization translates as "analytic", "arithmetic", and "mixed". However, these labels are changed here in order to align the terminology to international research.

still preserving the connection to the curriculum. A major disadvantage is that each individual decision is based on a much smaller dataset, as compared to a holistic judgment taking all available evidence about student proficiency into account[2]. Figure 1 provides an overview of the different models of grading (not including the intuitive model, which is presumably low in both validity and reliability) in relation to the alignment with the curriculum and the amount of data on student performance. As can be seen in the figure, the arithmetic model is low on alignment since it reduces the complexity of the data, which is done by transforming assessment outcomes to scores or marks that can be manipulated mathematically. This may result in higher reliability, but at the expense of validity. The holistic model preserves high detail in the data on student performance, potentially making grading more valid (depending on how the data is used), but is likely to result in low reliability. In the analytic model, on the other hand, each decision is based on a smaller amount of data (i.e. individual assignments), but each dataset still has a clear connection to the curriculum. This combination may result in at least moderately high validity and reliability.

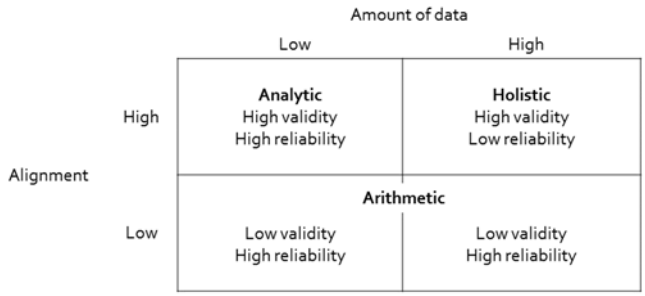

Figure 1. The different models of grading (not including the intuitive model) in relation to the amount and complexity of the data on student performance. Note that the fields represent theoretical predictions, which are not empirically tested.

Purpose and research questions

It is currently not known how the analytic model for grading compares to the holistic model in terms of validity and reliability. This study therefore aims to compare these models by investigating the agreement between teachers using the different models (reliability), as well as the justifications for their decisions (validity). Specifically, the study will answer the following research questions:

1. To what extent do teachers agree on students' grades when using an analytic or holistic model of grading?
2. How do teachers justify their decisions when using an analytic or holistic model of grading?

Methodology

The overall design of this study is experimental, where a number of teachers have been randomly assigned to two different conditions: analytic (n=13) or holistic (n=11) grading. Teachers volunteered to participate in the study and come from different schools in the same region. No personal data has been collected; only grades and written justifications from the teachers.

Procedure

In the analytic condition, the teachers received written responses to the same assignment from four students at four occasions during one semester (i.e. a total of 16 responses). The assignments all addressed writing in English as a foreign language (EFL), but otherwise had different foci (Table 1). All responses were from students aged 12, but with different proficiency in English. The responses were authentic responses from students, which had been anonymized.

Table 1. The four writing assignments.

Task n:o	*Assignment*
1	Writing a biography of a relative or a friend of the family
2	Writing an argumentative text about food waste
3	Writing a text about what a friend should be like
4	Writing a short text message (sms) to someone you care (or cared) about

The teachers were asked to grade each student response within a week after receiving them. In the end of the semester, teachers were asked to provide an overall grade for each of the four students, accompanied by a justification. This was done at a specific time and

[2] Another major disadvantage, for formative purposes, is that grades on individual assignment may have negative consequences for student learning (e.g. Black, Harrison, Lee, Marshall, & Wiliam, 2004).

place, with all the teachers, in order to standardize the procedure. The final grades and written justifications were used as data in the study.

In the holistic condition, participants were given the entire material at one 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. In both conditions, it took between 90-120 minutes to perform the grading and write down the justifications.

Agreement between teachers

A common method for estimating the agreement between different assessors[3] (i.e. inter-rater agreement) is by using correlation analysis. Depending on whether it is scores (a continuous variable) or grades (a discrete variable), either Pearson's or Spearman's correlation may be used. Spearman's correlation (ρ) 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 1 and E (i.e. the lowest passing grade) to 5. 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 (see Figure 2). In this study, therefore, Spearman's correlation has been combined with an estimation of absolute agreement in percent, as well as Cohen's κ, which takes into account the possibility of the agreement occurring by chance (for an in-depth discussion of different measures, see Stemler, 2004). In order to compare the agreement of several assessors, pair-wise comparisons has been made. Reported estimates of agreement are therefore calculated means from the pair-wise comparisons and the Mann-Whitney U test was used to test for statistical significance.

Case 1		Case 2	
Assessor 1	Assessor 2	Assessor 1	Assessor 2
A	A	A	C
A	A	A	C
C	C	C	E
E	E	E	F
A	A	A	C

Figure 2. In Case 1, both assessors agree on the exact grade for all students. The absolute agreement is 100% and the correlation is 1. In Case 2, the assessors agree on the rank order of students, but not on the exact grade since Assessor 1 is systematically more lenient than Assessor 2. In this case, the absolute agreement is 0%, while the correlation is still 1.

Justifications by teachers

The justifications for the grades, which were written down on paper, 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 recognize the full spectrum of teachers' language describing quality.

Second, all nodes were grouped in relation to six commonly used criteria for assessing writing (i.e. mechanics, grammar, organization, content, style, and voice). In addition, some teachers made references to comprehensibility and whether students followed instructions and finished the task. Some also made inferenced about students' abilities or willingness to communicate. Three additional criteria, called "Comprehensibility", "Rigor", and "Student", were therefore added to the categorization framework (Table 2). For an example of the categorization procedure, see Figure 3.

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

[3]This discussion applies equally well to intra-rater agreement (i.e. the agreement between assessments made by the same assessor, but at different occasions). Since this is not part of the current investigation, however, intra-rater agreement is not further discussed.

Table 2. Criteria for assessing writing used for categorization.

Criteria	*Description*
Mechanics	Use of accurate spelling and punctuation
Grammar	Use of appropriate grammar and standard English
Organization	Organization, structure, and use of strategies to aid in comprehension
Content	Level of detail, use of comparisons, examples, and arguments
Style	Appropriate use of words, sentences, and paragraphs; flow and variety
Voice	Personality and sense of audience
Comprehensibility	Whether the text is understandable to the reader
Rigor	Adherence to instructions, doing revisions, and finishing the task
Student	Ability of the student and the willingness to communicate

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 [Organization]. 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 discernable [Voice] and can be understood by an English-speaking person [Comprehensibility]. (Justification for Student 1; Teacher 2 in analytic condition)

Figure 3. Typical example of justification and categorization (in square brackets).

Findings

Agreement

Thirteen teachers participated in the analytic condition, which means 91 pair-wise comparisons. The mean agreement in this group was 66.2 percent, which means that the teachers agreed on the exact same grade in two thirds of the cases, with a standard deviation of 21.2. The mean rank correlation was .973.

In the holistic condition, there were 11 teachers and 55 pair-wise comparisons. The mean agreement in this group was 45.9 percent, which means that the teachers agreed on the exact same grade in about half of the cases, with a standard deviation of 23.0. The mean rank correlation was .943. The statistics are summarized in Table 3.

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

	Analytic condition	*Holistic condition*
Percent agreement (Std. deviation)	66.2 (21.1)	45.9 (23.0)
Cohen's κ	.602	.405
Spearman's ρ (Std. deviation)	.973 (.026)	.943 (.064)

As can be seen by comparing the statistics from the different conditions, the mean agreement for the analytic condition is considerably higher as compared to the holistic condition, and the standard deviation is also somewhat smaller. The correlation is slightly higher in the analytic condition and the standard deviation is smaller. Although the correlation is relatively high (i.e. above .9) in both conditions, the difference is still statistically significant at the $p<.001$ level.

Justifications

All in all, the teachers in the sample made 537 references to quality indicators in their justifications (i.e. on the average 22.4 references per teacher), using 64 different terms for describing these qualities. As can be seen in Table 4, although the teachers in the holistic condition made slightly more references, the difference is quite small (on average one reference more per teacher and student).

Table 4. Overview of justifications for conditions and students.

	Student 1	*Student 2*	*Student 3*	*Student 4*	*Sum*	*Mean*
Analytic	66	79	63	66	274	19,6
Holistic	63	79	55	66	263	23,9
Total	129	158	118	132	537	22,4

Table 5 summarizes teachers' references in relation to the criteria. 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.

There are some differences between the conditions, most notably that teachers in the holistic group provided more references to organization. This group also made more references to rigor and inferences about the students, but the number of references in these categories is comparably small.

In relation to the students, it was almost exclusively the justifications for the highest grade (i.e. Student 2) that included (positive) references to content and voice (and to some extent Student 4). On the contrary, it was the justifications for lower grades that made reference to (lack of) comprehensibility.

Discussion

This study aimed to compare the analytic and holistic models of grading by investigating the agreement between teachers using the different models, as well as the justifications for their decisions.

Comparing the two conditions

The statistical comparison shows that there is indeed a significant difference between the two conditions. The correlation analysis shows that teachers in the analytic condition have a higher correlation (including lower standard deviation) between the grades they assigned to the student responses. In addition, the absolute agreement is considerably higher. While the teachers in the holistic condition agree in less than half of the cases, the teachers in the analytic condition agree in approximately two thirds of the cases. Of course, in terms of comparability in grading, this may still not be considered acceptable. However, in relation to 45 percent agreement, it is nonetheless a substantial improvement.[4] On the contrary, the comparison of teachers' justifications suggests that there are no substantial differences between the conditions. Teachers in both groups provided approximately the same amount of references both within and across the different criteria. An exception is Organization, where teachers in the holistic condition provided significantly more references as compared to teachers in the analytic condition. It is difficult to explain this finding, since there are no major differences with regard to the other criteria. For instance, it could be hypothesized that teachers in the holistic condition would focus on surface features (such as organization, mechanics, and grammar), given that they had not had the opportunity to familiarize themselves with the tasks before the grading. However, there are no differences with regard to mechanics or grammar, and both conditions provided

Table 5. Summary of teachers' references in relation to the criteria for students and conditions.

	Student 1	*Student 2*	*Student 3*	*Student 4*	*Analytic (mean)*	*Holistic (mean)*	*Sum*
Mechanics	8	14	5	6	14 (1.0)	19 (1.7)	33
Grammar	15	19	19	17	43 (3.1)	27 (2.7)	70
Organization	18	18	12	22	29 (2.1)	41 (3.7)	70
Content	2	11	-	5	7 (0.5)	11 (1.0)	18
Style	53	67	48	56	117 (8.4)	107 (9.7)	224
Voice	4	19	4	7	20 (1.4)	14 (1.3)	34
Comprehensibility	20	2	19	7	29 (2.1)	19 (1.7)	48
Rigor	4	6	6	8	10 (0.7)	14 (1.3)	24
Student	5	2	5	4	5 (0.4)	11 (1.0)	16
Total	129	158	118	132	274 (21.1)	263 (23.9)	537

[4] Note that the agreement is influenced by the length of the grading scale. If using "adjacent agreement" (i.e. allowing for +/- 1 on the grading scale) instead of "exact agreement", thereby making the grading scale shorter, there would have been a 94 and 90 percent agreement in the analytic and the holistic conditions respectively.

the same amount of references in relation to less obvious criteria, such as style and voice.

Similar to the situation with the criteria, there are no differences between the groups in relation to the students. Rather, there is quite a strong consensus about the qualities in students' performances. For instance, both groups agree that Student 1 has a simple language, which is mostly comprehensible, although there are a number of disturbing grammatical errors. Similarly, both groups agree that Student 2 has a varied and well-developed language, and that the texts are well adapted to the purpose and audience. Among other things, 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 assessment may not necessarily be affected by the inequality inherent to grading, since 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 to what extent these criteria are fulfilled in students' texts. The teachers also agree on the rank order of student performance to a high extent. However, when assigning specific grades, the absolute 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.

Validity of the analytic model

The findings are in line with the model presented in Figure 1, where it is assumed that the analytic condition would result in at least moderately high validity and reliability, due to a reduction of complexity in the grading process. The holistic condition, on the one hand, was assumed to result in moderately high validity (similar to the analytic condition), but lower reliability, which was also the case. If striving towards higher agreement between teachers' grading, reducing complexity by adhering to an analytic grading model may therefore be a viable option. Such as strategy, however, may be considered in conflict with "interpretivist approaches" (e.g. Moss, Girard, & Haniford, 2006), which stress the importance of holistic integration of available sources, as opposed to a more selective sampling of data supporting the initial (supposedly intuitive) interpretations (Nijveldt, 2007). Still, in the current study the teachers' written justifications were based on shared criteria and there were no indications of justifications from teachers in the holistic condition being different from teachers in the analytic condition.

That the justifications from the teachers were similar in both conditions does not, of course, guarantee the validity of the grading process. Although the current study cannot identify which factors (beyond the criteria) that influenced teachers' grading, since these factors were not present in teachers' written justifications, the agreement between teachers' grades is still generally low and a lot of variance is left unexplained. This variation may be due to teachers attaching different weight to different criteria, but could also be explained by individual preferences and contextual factors (e.g. Kunnath, 2017; McMillan, 2003). Consequently, there is room for improvement and educating teachers in using strategies for considering and combining evidence, as well as addressing potential threats to validity (Nijveldt, Beijaard, Brekelmans, Wubbels, & Verloop, 2009), may very well support such improvements, but this needs to be further investigated.

Tentative conclusions and implications for practice

The findings from this study suggest that analytic grading, where teachers assign grades to individual assignments, and use these "assignment-grades" when deciding on the final grade, is preferable to holistic grading in terms of reliability. Teachers from both groups were in agreement on which criteria to use when assessing student work and the qualities identified in students' performance, which means that there is no reason to believe that the conditions would differ in terms of validity.

In should be noted, however, that this practice may not necessarily be optimal for formative assessment. Therefore teachers may consider keeping the "assignment-grades", which are based on limited data and assumingly unreliable, to themselves, while only communicating qualitative feedback to the students (i.e. strengths and suggestions for improvements) in order to support students' learning and improved performance.

Recommendations for future research

The main contribution of this study lies in the assumptions tested (Figure 1), which have the potential to explain the difference in agreement between different models of grading. The same methodology can therefore be used to test the same assumptions, but with a larger and more representative sample of teachers; both in EFL (as here) and in other subjects.

Furthermore, the grading process in this study included written performance only, which means that a more heterogeneous material, including – for example – oral performance, would have provided a more valid comparison with teachers' actual grading practices. However, a more heterogeneous material could also be assumed to further accentuate the differences between the models of grading, by making the holistic grading even more complex, possibly resulting in lower agreement.

As mentioned above, the current study cannot confirm which individual and contextual factors that influence teachers grading, only that some factors beyond the grading model – such as giving different weight to different criteria – give rise to variability in the sample. Given the great variability of assigned grades, as well as the fact that this influence has been a robust finding in numerous studies across 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 judgments to what they believe are legitimate criteria, since they know that someone will evaluate their assessments.

Taken together, it is recommended that the findings from this study are further investigated by using a larger sample of teachers, a more heterogeneous material, and by including other subject areas. It is also recommended to investigate to what extent educating teachers in using strategies for considering and combining evidence, as well as addressing potential threats to validity, may support the validity and/or reliability of grading.

References

Black, P., Harrison, C., Lee, C., Marshall, B. & Wiliam, D. (2004). Working inside the black box: *Assessment for learning in the classroom*. Phi Delta Kappan, 86, 8-21.

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

Brookhart, S. M. (2013). The use of teacher judgement for summative assessment in the USA. *Assessment in Education: Principles, Policy & Practice*, 20, 69-90.

Brookhart, S. M., & Chen, F. (2015). The quality and effectiveness of descriptive rubrics. *Educational Reviews*, 67, 343–368.

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.

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

Eells, W. C. (1930). Reliability of repeated grading of essay type examinations. *Journal of Educational Psychology*, 21, 48-52.

Gipps, C., Brown, M., McCallum, B. & McAlister, S. (1995). Intuition or Evidence? Buckingham: Open University Press.

Isnawati, I., & Saukah, A. (2017). Teachers' grading decision making. *TEFLIN Journal,* 28, 155-169.

Jonsson, A. & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, 2, 130-144.

Kilday, C. R., Kinzie, M. B., Mashburn, A. J., & Whittaker, J. V. (2012). Accuracy of teacher judgments of preschoolers' math skills. *Journal of Psychoeducational Assessment*, 30, 148-159.

Klapp, A. (2015). Does grading affect educational attainment? A longitudinal study. *Assessment in Education: Principles, Policy & Practice*, 22, 302-323.

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, Malmo: Malmo University.

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.

McMillan, J. H. (2003). Understanding and improving teachers' classroom assessment decision making: Implications for theory and practice. *Educational Measurement: Issues and Practice*, 22, 34-43.

McMillan, J. Myran, S., & Workman, D. (2002). Elementary teachers' classroom assessment and grading practices. *Journal of Educational Research*, 95, 203-213.

Moss, P.A., Girard, B.J., & Haniford, L.C. (2006). Validity in Educational Assessment. *Review of Research in Education*, 30, 109-162.

Nijveldt, M. J. (2007). Validity in Teacher Assessment. An Exploration of the Judgement Processes of Assessors. Doctoral dissertation, Leiden: Leiden University.

Nijveldt, M., Beijaard, D., Brekelmans, M., Wubbels, T., & Verloop, N. (2009). Assessors' perceptions of their judgement processes: Successful strategies and threats underlying valid assessment of student teachers. *Studies in Educational Evaluation*, 35, 29-36.

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

Randall, J., & Engelhard, G. (2008), Differences between teachers' grading practices in elementary and middle schools. *The Journal of Educational Research*, 102, 175-185.

Randall, J., & Engelhard, G. (2010). Examining the grading practices of teachers. *Teaching and Teacher Education*, 26, 1372-1380.

Sadler, R. D. (1987). Specifying and promulgating achievement standards. *Oxford Review of Education*, 13, 191-209.

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

Starch, D., & Elliott, E. C. (1913a). Reliability of grading work in mathematics. *The School Review*, 21, 254-259.

Starch, D., & Elliott, E. C. (1913b). Reliability of grading work in history. *The School Review*, 21, 676-681.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, 9.

Citation:

Jönsson, Anders, & Balan, Andreia. (2018). Analytic or Holistic: A Study of Agreement Between Different Grading Models. *Practical Assessment, Research & Evaluation*, 23(12). Available online: http://pareonline.net/getvn.asp?v=23&n=12

Corresponding Author

Anders Jönsson
Professor of Education
Kristianstad University
Sweden

email: anders.jonsson [at] hkr.se