*Globe: A Journal of Language, Culture and Communication*, 9: 133-154 (2020)

# Adaptive grading systems, or pros and cons of different ways of grading grammar exams

*Richard Skultety Madsen, Aalborg University*

**Abstract:** This paper investigates several alternatives to the grading system used currently when examining students' knowledge of theoretical grammar in the Department of English Business Communication at Aalborg University, Denmark. The proposed alternatives differ from the current system in two parameters, namely by differentiating between exam questions according to their levels of difficulty and by evening out biases which are due to the differences in the weights of the various topics of the exam. It is found that the proposed methods would yield results significantly different from the current grading method even though it would only be in the favor of few students in terms of better grades to adapt any of them. Nevertheless, the study reveals prevalent traits of the current way of examining, such as built-in bias and the scalability of the questions, which are important considerations to anyone conducting exams, not just in grammar. Furthermore, the paper uncovers unexpected features of clause constituents that may have serious implications for their teaching.

**Keywords:** language acquisition, learning of grammar, evaluating and grading, statistics.

## 1. Introduction

The purpose of this paper is to investigate how the grading of an exam can be fine-tuned. The investigation is based on the exam in theoretical grammar of freshmen in the Department of English Business Communication at Aalborg University, Denmark. However, the methods tested can be adapted to any exam or test which is graded quantitatively; that is, the students are given a certain number of points for each exam question answered, and the grade is then dependent on the sum of the points so collected.

The idea for this study came from a project planned previously to correlate students' vocabulary as manifested in their written assignments with their grammatical knowledge as measured by the grammar exams. During the planning of said project, it was realized that the scores of the grammar exams would not be able to differentiate the students sufficiently. In the current exam scheme, each of the 100 questions answered correctly is awarded one point. Thus, the exam scores can differentiate at most 101 students (0 thru 100 points); in practice even fewer students because not all possible scores are actually attained (very few students score above 95, and virtually no one has ever scored under 40). This would have made the correlation analysis less useable.

Therefore, a method was sought that could retroactively increase the granularity of the grammar exams which had already been administered. Even though the vocabulary-correlated-with-grammar project has not been pursued further yet, it was thought that methods to increase the granularity of the exam scores would nonetheless be worth investigating in their own right with a view to refining the way of examining the students in grammar without having to change the examination fundamentally. A good reason for possibly changing the evaluation process is that there is currently no differentiation between the exam questions reward-wise even though it is likely that they represent different levels of difficulty. Thus, it is possible that a student who is able to answer only less difficult questions scores more points and consequently a higher grade than a student who is able to answer fewer but more difficult questions. This is a potentially unfair or undesirable situation.

This paper draws up several ways of differentiating between the exam questions and investigates the consequences of these methods by comparing them to the current manner of examining. It does not address general issues surrounding grading, such as the reasons for or the goals

with grading (Brookhart 2011; Aitken 2016). It restricts itself to fine-tuning the current way of examining. As an extension to devising new scoring methods, this study also tests whether there might be unwanted biases in the current method of examining in theoretical grammar. An early draft of the paper was presented at a departmental seminar in 2017.

## 2. Theory

This study focuses on how it is possible to take the difficulty level of the exam questions into consideration in order to fine-tune the grading process. The methods that are examined here are based on assigning to each question a different number of points in accordance with the level of difficulty of the questions. Hence, determining the level of difficulty of each question is of crucial importance. The grading system itself is not modified; that is, the relative distances between the grades are not changed (Ministry of Education 2019). The grading system is kept intact so that it is easier to determine the consequences of the fine-tuning methods investigated.

There are in principle two ways in which the level of difficulty can be set, a priori and a posteriori. In the a priori approach, the level of difficulty – in the form of different number of points that can be scored by answering the questions correctly – is assigned to each question before the exam is attempted. In the a posteriori approach, the level of difficulty of the questions is calculated after the exam has been attempted by the students. This paper follows the a posteriori approach. In the remainder of this section, it is explained why not the a priori method is favored, and the section on methods elaborates which a posteriori methods are investigated and how they are implemented.

There are two reasons why the a priori approach is not used in this paper. One reason is simply that this paper compares different ways of grading on the basis of an exam that has been taken and whose questions had not been differentiated with respect to their difficulty. The other reason is that it is in fact non-trivial to assess the level of difficulty of questions beforehand even if – intuitively – it should be the preferred method. There are basically two ways of doing it.

One way is the intuition of the examiner. All teachers/examiners develop a feeling of what tends to be more and what tends to be less difficult for the students, and this intuition is likely drawn upon when selecting the questions for an exam. However, the problem is that it is only an intuition. There ought to be another, more scientific way of performing the selection of exam questions, especially when the examiner is the same person as the teacher, and when this person is almost alone in this process (Lehmann 2018). This is certainly the case in our Department of English Business Communication, as there is no tradition in Denmark to have centrally standardized exams at universities, and there are currently only two teachers who teach grammar. Thus, great responsibility rests upon the examiner in order to avoid bias and keep the level of the exam as constant as possible across the years.

Another, more objective, way of assessing the difficulty of the questions is to analyze the responses of students at previous exams and assign points to the questions of future exams based on this statistical study. Unfortunately, there are two problems with this approach.

One of the challenges is that the questions – of course – have to be different from exam to exam, or else the students of later years would have a great advantage compared to the students of the first year. Hence, the assignment of points to the individual questions of a future exam would be dependent on the extent of analogy between the new questions and the questions that have been assessed in the above-mentioned statistical analysis. However, the extent of analogy itself could only be assessed by either the examiner's intuition or by an even more extensive statistical analysis that takes many different types of questions across the years into account.

However, the major challenge to this approach is that such statistical analyses simply do not yet exist, at least not for the types of questions posed at the grammar exams in our department. Madsen (2017) is a fairly large scale statistical study of our students' performance at the grammar exams; however, it focuses on the question to what extent the different topics of grammar challenge the

students. It does not assess the level of difficulty of individual grammar-exam questions. Another study (Madsen ms) does examine the questions individually. However, its focus lies elsewhere and therefore also considers questions in the exercises which the students do during the grammar course as part of their preparation for the exam.

Based on the above considerations, it seems that the most promising approach – at least for the time being – is the a posteriori methodology, which is elaborated in the section on the methods. Since there are no specific expectations to the outcome of this study, no hypotheses are postulated. Hence, the study is predominantly inductive. Therefore, the data are presented in the next section before the methods are discussed.

Of course, the methods investigated here do not guarantee that the exams in different years have the same overall level of difficulty, reliability and validity, nor do they it ascertain that the differences between the levels of difficulty of questions found reflect a tendency in the population, i.e. outside the sample of students. However, it is not the purpose of these methods, either. Their purpose is to make the assessment of the exam scores fairer. On the other hand, this study does serve as a step in investigating whether the difficulty of exam questions is implicational or not.

There is an intuitive expectation that if someone can manage a task considered more difficult (on whatever basis), they can also manage less difficult tasks, but not vice versa (Vygotsky 1978; Hatch & Farhady 1982; Donato 1994). This is an implicational relation. The ability of doing something more difficult implies the ability of doing something less difficult, but not the other way around. However, it need not be the case. For instance, vocational educations are considered to be at a lower level than university educations, suggesting that they are easier (Ministry of Education 2018). Nevertheless, it does not guarantee that say a person with a PhD could take on a plumber's job. By evaluating the methods which are proposed here for a more differentiated grading, the implicationality of the exam questions is analyzed as well.

## 3. Data

The data that are manipulated according to different methods are the students' scores from the grammar exam in 2014. It is a written exam in theoretical grammar; that is, the students' practical command of English is not tested apart from 5 questions concerning the use of comma in certain sentences. The exam consists of 100 questions on 13 topics. The students are given 120 minutes to answer the 100 questions and are not allowed to use any means of aid. Consequently, they have to memorize all the relevant technical terms and their applicability. Table 1 gives an overview of the topics of grammar in the exam.

Table 1: Overview of the grammar topics examined

| Topics | Number of questions |
|---|---|
| Parts of speech | 10 |
| Semantic relations | 5 |
| Clause constituents | 18 |
| Phrase vs. subordinate clause | 8 |
| Phrase types | 10 |
| Phrase constituents | 9 |
| Pronoun types | 10 |

| Topics | Number of questions |
|---|---|
| Subordinate clause types | 7 |
| Clause finiteness | 7 |
| Number of matrix clauses in a paragraph | 5 |
| Function of a morpheme | 3 |
| Dictionary form of a word's root | 3 |
| Comma | 5 |

The numbers of questions per topic, which might seem ad hoc, are the result of a compromise between four factors. The period of two hours allotted to the exam is decided externally and sets the limit for how many questions overall it is reasonable to pose. On the other hand, as many topics as possible are probed for the sake of the validity of the exam. Then, reliability requires that as many questions as possible are asked per topic (DeVellis 2011; Dörnyei 2014), and preferably, about the same number of questions per topic so that there is as little bias as possible towards select topic(s). Finally, tradition also plays a role, as for instance, clause constituents used to be highly represented whereas morphology did not use to be represented at all in previous exams. Table 2 provides some examples of the questions posed within the different topics.

Table 2: Examples of exam questions

**Determine which part of speech the underlined words belong to.**

- The name Intel is a portmanteau of **Int**egrated **El**ectronics.

**Determine the semantic relation between the expressions below.**

- *-er* as in *happier* vs *-er* as in *Londoner*

**Determine what clause constituents the underlined sequences of words are.**

- True cider is made from fermented apple juice.

**Decide whether the underlined sequences of words are phrases or clauses.**

- Founded in 1968, Intel mostly produced RAM in the beginning.

**Determine what phrase constituent the underlined sequences of words are.**

- the transistor count of modern processors

**Determine what kind of pronoun the underlined words are.**

- Not even Intel itself has anticipated its success.

**Determine the type and finiteness of the underlined subclauses.**

- It seems that some drinks marketed as cider are not true ciders.

**Specify the dictionary form of the roots of the words below.**

- Unhealthily

With the exception of the questions in which the students have to provide the dictionary forms of the roots of words type, the students have to select the correct answer from finite sets of valid answers. For instance, in the case of clause constituents, the set of valid answers is the set of clause constituents, containing nine elements in this grammar course, such as subject, verb, direct object, indirect object, subject complement, object complement, adverbial constituent, preliminary subject and preliminary direct object (Hjulmand & Schwarz 2008). Should a student give a true but invalid response, say calling *from fermented apple juice* a preposition phrase instead of an adverbial constituent, the response counts as incorrect. The sets of valid responses are not listed in the exam; the students are expected to remember them. Hence, the exam is not a classic multiple-choice exam. In questions concerning the roots, there is no fixed set of valid responses, and the students are not given any hints as to what the root might be.

Each correct and valid response yields one point for the student. Incorrect and non-existent responses yield zero points. The students have to collect 60 points (60% of the maximum number of points) in order to pass the exam. The boundaries for the grades can be seen in Table 3 (Ministry of Education 2017). There is no provision for partially answered questions. Hence, fractions of points are not given. In any case, only the questions concerning semantic relations, the roots of words and the use of comma could conceivably be answered partially in a meaningful way, for instance if a student inserts only one comma into a sentence that requires two commas.

Table 3: Grade boundaries

| Grade | Boundaries |
|---|---|
| -3 | 0 – 17 |
| 00 | 18 – 59 |
| 02 | 60 – 63 |
| 4 | 64 – 73 |
| 7 | 74 – 85 |
| 10 | 86 – 95 |
| 12 | 96 – 100 |

## 4. Method

This section explains both the new methods of grading and the method used for measuring the implicationality in the perceived difficulty of the exam questions.

### *4.1. Proposed grading methods*

An important consideration for the grading methods to be investigated is that they can be integrated seamlessly into the process in which the exams are conducted currently. Thus, the examiner should not have to do anything else than deciding whether a question has been answered correctly or not. The methods have been implemented in a MS Excel spreadsheet and require nothing else than entering 1 for a correct answer and 0 for an incorrect answer (Bovey et al. 2009; Carlberg 2014). It was contemplated whether non-existent replies and/or invalid responses should be treated in a special manner. However, since there is no tradition for penalizing the students for such responses, they are treated simply as incorrect answers and are thus to be assigned 0.

Generally, an a posteriori assessment of the level of difficulty can be performed by calculating the ratio of how many students have answered the questions correctly (Hatch & Farhady 1982: 177). If the sample size is large enough, in the present case 68 students, this figure can be expected to reliably indicate the level of difficulty of the questions relative to each other. The higher the ratio of correct answers, the easier the question. The questions are only compared to other questions within the same topic. A cross-topic comparison of individual questions would not make much sense as oranges would be compared to apples, especially since it has been established that some topics are generally more difficult than others (Madsen 2017).

A consequence of a posteriori assessments is that the students cannot be informed beforehand which questions are considered more difficult and hence yield more points. Another consequence is that the calculation is specific for the group of students who take the exam together, and once their grades have been fixed, the group cannot be expanded because the grades of the students depend on each other. However, this is not likely to ever be an issue in practice.

Two methods have been devised to differentiate between the questions regarding their relative levels of difficulty. They differ with respect to how the differentiation is done. Assuming that n equals the number of questions within a given exam topic, the one method assigns an integer ranking value from 1 thru n to each question depending on the detected level of difficulty. 1 indicates the lowest level of difficulty, i.e. the highest number of informants having answered that question correctly. If two questions appear to have the same level of difficulty, i.e. they have been answered correctly by the same number of informants, they are assigned the same value. N indicates the highest level of difficulty.

The score of a given student is computed by first adding the number of correctly answered questions to the sum of the ranking values of the correctly answered questions and then divided by a divisor specific to the given topic. Suppose a student answers the questions in a topic with the ranks 1, 3, 6 and 7 correctly out of 10 questions. Then their raw score is (4 + 1 + 3 + 6 + 7) / 65 = 0.323. The raw score is always a value between 0 and 1, both inclusive. The divisor, in this case 65, represents the maximum granularity of the score for the given topic and derives from the number of questions (n) and equals n + n * (n+1) / 2. Granularity represents the maximum number of distinct values that can be distinguished within the given topic; that is, the maximum number of students that can be differentiated from one another based on their responses. Without this kind of question differentiation, 10 questions can only differentiate between 11 students as there are only 11 possible, numerically different outcomes (0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 points). The increase in granularity is especially useful in the case of topics with few questions. For instance, in the case of 5 questions, the maximum granularity thus achieved is 20, which is much higher than 6 in the case of the undifferentiated score (0, 1, 2, 3, 4, 5 points). However, if there are questions with the same level of difficulty, granularity decreases, but only in the worst case scenario – all the questions having the same level of difficulty, which is highly unlikely – does it fall to the level of granularity of the undifferentiated scores.

By adding the number of correctly answered questions (4 in the example above), it is possible to distinguish between situations that would otherwise yield the same score. Suppose one student answers the question with rank 3 correctly, and another student answers the questions with ranks 1 and 2 correctly. Without including the number of correctly answered questions, they would both gain 3 / 55 = 0.0545 raw points[1]. However, by using the above formula, the first student gains 4 / 65 = 0.0615 raw points, and the second one 5 / 65 = 0.0769 raw points.

The logic behind assigning a higher score to the second student is twofold. At a philosophical level, broader knowledge is purposefully valued higher even if it only manifests itself in the ability

[1] The reason why the divisor is 55 here is that if the number of questions is not taken into account, the divisor is only 55, not 65, in the case of 10 questions.

of answering easier questions correctly. At a statistical level, it is more likely that one answers one (or generally fewer) question correctly by pure chance than two (or generally more) questions. Hence, this way of calculating the score purposefully favors those who answer more questions correctly because it is less likely to happen by pure luck. Moreover, also taking the number of questions answered correctly into account has the intended effect that only those informants who answer questions with the exact same levels of difficulty correctly gain the same score.

The other method of differentiating between the questions works in the same way as the one just described except the fact that the level of difficulty is not expressed by integer, but by rational numbers. This has the effect that the calculation takes into account not only the rank order of the questions on the scale of difficulty, but also the proportion of how difficult they are compared to one another within the same topic. The reason is that the difficulty of the questions need not be equally distributed. For instance, the second easiest question may be five times more difficult than the easiest question, and the third easiest question only slightly more difficult than the second easiest one – as measured in terms of the number of informants having answered them correctly.

This distributional difference is disregarded in the first type of differentiation, explained above. In the method of proportional differentiation, the calculation is done in the following manner. The easiest question is assigned the value of 1, and the most difficult question is assigned the value of the number of questions in the given topic, say 10. All the other questions are assigned values between 1 and the highest value (say 10) proportionally to their measured difficulty. One concrete example, pertaining to the topic of parts of speech, indicating how uneven the distribution of difficulty may be: 1, 1.6, 1.75, 1.9, 2.8, 2.8, 4, 4.9, 7.9, 10.

In this method, the divisor (that is, the maximum number of points that can be achieved in a given topic) cannot be calculated by a formula on the basis of only the number of questions as in the method of ordinal differentiation above, but only by adding the actual ranks together, in the example 48.65 = 1 + 1.6 + 1.75 + 1.9 + 2.8 + 2.8 + 4 + 4.9 + 7.9 + 10 + 10 (number of questions). This means that even topics with the same number of questions can have different divisors. However, it does not influence granularity, which is always n + n * (n+1) / 2, only the weighting of the individual questions varies. In all other respects, the two methods of differentiating between the questions work in the same way.

The divisors in the proportional method mostly happen to be lower (as in the example above) than the corresponding ones in the ordinal method, but are sometimes higher. It depends on the actual distribution of the difficulty levels. The fact that the divisors in the proportional method tend to be lower than the corresponding divisors in the ordinal method suggests that the distribution of difficulty within the topics is biased towards the lower end. I.e. most of the questions tend to be relatively easy while one or two questions are exceptionally difficult compared to the other ones.

Since the raw scores gained from the differentiation methods are not directly comparable to the scores gained by simply counting the questions answered correctly, they must be scaled up. In this paper, two methods of scaling have been tried. In one of the methods, the raw scores are scaled up to be in the same range as the scores gained from just counting the correct answers. In the example above, it is from 0 thru 10. This is achieved by multiplying the raw score by 10, or generally by the number of questions (n). Hence, if a student answers the questions in a topic with the ranks 1, 3, 6 and 7 correctly out of 10 questions, his score is 3.23. This score happens to be lower than what the student is given under the traditional scheme (4 points); however, it is precisely the idea of the differentiation between questions that students gain different scores depending on which questions they answer. This method of scaling up makes it possible to calculate the grades in the exact same manner as usual since the range of the aggregate score (the sum of the scores in the individual topics) remains the same, namely 0 thru 100 in our grammar exam.

The other method of scaling up transforms the raw scores into percentages, which is achieved by multiplying the raw scores uniformly by 100 instead of the number of questions concerning the

given topic. This method has the advantage that it makes it possible to compare the students' performance regarding the individual topics. It also evens out the built-in bias between the topics, which derives from their being probed by different numbers of questions. On the other hand, it has the slight disadvantage that the calculation of the grades has to be modified because the range of the aggregate scores changes. It will be 0 thru 1300% in the case of 13 topics in the exam. However, this modification of the grading is rather simple since the boundaries of the grades are defined in terms of percentages of the maximum attainable score. Thus, 1300(%) simply has to be set as the maximum score, and the grade boundaries have to be recalculated from it according to the values in Table 3. Alternatively, the percent points have to be divided by 13 (there being 13 topics in the exam) to bring the scores in percent points into the same range as the scores according to the original grading method.

To sum it up, 6 ways of evaluating the grammar exam will be compared as shown in Table 4.

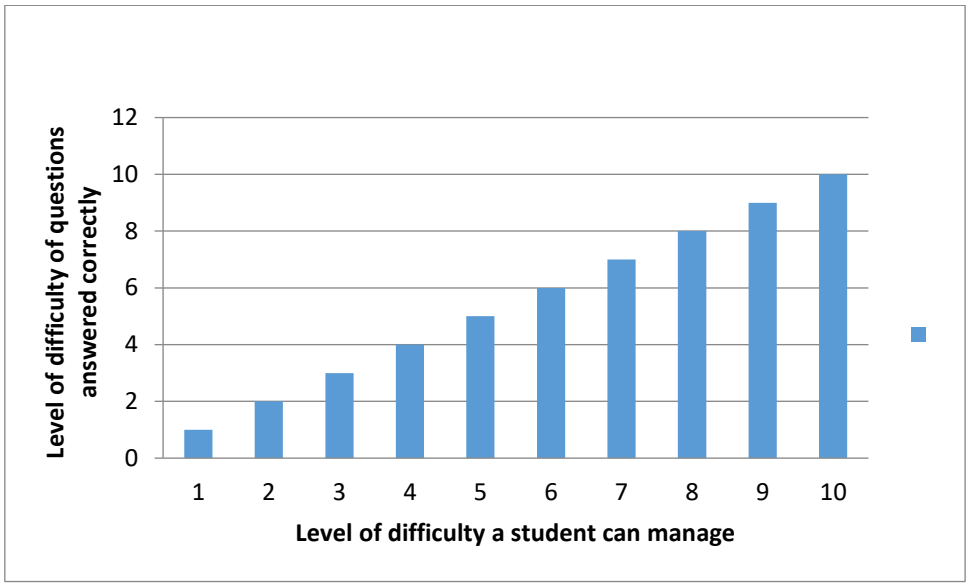
Table 4: Overview of grading methods

| | Description | Name henceforth |
|---|---|---|
| 1 | Counting the questions answered correctly for the exam question set as a whole (the original method) | Original |
| 2 | Counting the questions answered correctly separately for each topic and transforming these subscores into percentages, which are summed (evening out the representational bias between the topics) | % |
| 3 | Differentiating between the questions within each topic ordinally and then adding up these subscores | ordinal |
| 4 | Differentiating between the questions within each topic ordinally and then transforming these subscores into percentages, which are summed (evening out the representational bias between the topics) | ordinal% |
| 5 | Differentiating between the questions within each topic proportionally and then adding up these subscores | prop |
| 6 | Differentiating between the questions within each topic proportionally and then transforming these subscores into percentages, which are summed (evening out the representational bias between the topics) | prop% |

*4.2. Measuring implicationality*

Assuming that the difficulty of the questions is implicational, perfect implicationality or scalability (Hatch & Farhady 1982; Bettoni & Di Biase 2015) can be depicted as in Figure 1.

Figure 1: Perfect implicationality of difficulty




The diagram indicates that students must be able to answer all the questions which are below the highest level of difficulty they can manage, but none of the questions which are above it. For instance, if a student is able to manage say difficulty level 6, he or she must be able to answer all the questions of difficulty levels 1 thru 5, but none of difficulty levels 7 thru 10. The method used to calculate implicationality orders the questions in the individual topics according to their level of difficulty as indicated by the number of students able to answer them correctly. Then, in descending order of level of difficulty, it counts how many other questions, i.e. questions below the maximum level of difficulty managed by them, the students have answered. Since it is only possible to measure the students' level of knowledge from their answers to the exam questions, there is no way of knowing whether a student who say only answered questions 1 thru 5 correctly might have been able to answer say question 7 correctly, too, but "just" missed it at the exam. This is why the method takes the question with the highest level of difficulty answered correctly as the starting point and only counts downwards. No attempt is made to distinguish between the lower-level questions. That is, if the starting point is say level 7, then the two sequences of questions answered correctly 1, 2, 3, 4, 5 and 2, 3, 4, 5, 6 are treated as equal. Both sequences count as having missed one lower-level question. Finally, the averages of these counts for each starting level of difficulty are computed. If there is substantial linear implicationality in the difficulty levels, the resulting averages should plot a diagram similar to the one in Figure 1.

## 5. Analysis

Table 5 shows the distribution of grades gained from the different methods. The averages of the % methods have been brought into the same range as that of non-% methods for the sake of comparability.

Table 5: Overview of grades from the various grading methods

| Grades | Original | Ordinal | Prop | % | Ordinal% | Prop% |
|---|---|---|---|---|---|---|
| **12** | 0 | 0 | 0 | 0 | 0 | 0 |
| **10** | 3 | 0 | 0 | 1 | 0 | 0 |
| **7** | 26 | 15 | 14 | 21 | 12 | 11 |
| **4** | 23 | 17 | 16 | 23 | 19 | 16 |
| **02** | 5 | 9 | 11 | 8 | 6 | 9 |
| **00** | 11 | 27 | 27 | 15 | 31 | 32 |
| **-3** | 0 | 0 | 0 | 0 | 0 | 0 |
| **Avg. grade** | 4.62 | 2.81 | 2.71 | 3.90 | 2.53 | 2.34 |
| **Avg. points** | 69.7 | 63.0 | 62.0 | 67.6 | 61.2 | 60.1 |

As can be seen, none of the new methods are in the students' favor generally. Only one single student would have gained a higher grade, namely 4 instead of 02 had the exams been graded according to the % method, which is the method closest to the original one. In all other cases, the students would receive the same or lower grades compared to the original method.

Interestingly, all the methods that attempt to even out the bias in the contribution of the individual topics to the final grade (the % methods) yield lower grades than the corresponding biased methods. This fact suggests that those topics which are probed by the most questions and thus have the strongest weights happen to be the ones that contribute positively to the students' grades, i.e. the students tend to be somewhat better-versed in them than in the topics which are tested by fewer questions. This is especially true of the morphological topics, in which most students do not do well, which, however, gain weight in the % method compared to the original method. Table 6 shows the difference between the original grading method and its % counterpart with respect to the weights of the topics.

Table 6: The weights of the individual topics

| Topics | Avg. scores | Weight in original method | Avg. contribution to the total number of points in the original method | Weight in % method | Avg. contribution to the total number of points in the % method |
|---|---|---|---|---|---|
| Parts of speech | 0.646 | 10 | 6.46 | 7.69 | 4.97 |
| Semantic relations | 0.694 | 5 | 3.47 | 7.69 | 5.34 |
| Clause constituents | 0.641 | 18 | 11.53 | 7.69 | 4.93 |

| Topics | Avg. scores | Weight in original method | Avg. contribution to the total number of points in the original method | Weight in % method | Avg. contribution to the total number of points in the % method |
|---|---|---|---|---|---|
| Phrase vs. subordinate clause | 0.739 | 8 | 5.91 | 7.69 | 5.68 |
| Phrase types | 0.846 | 10 | 8.46 | 7.69 | 6.50 |
| Phrase constituents | 0.683 | 9 | 6.15 | 7.69 | 5.25 |
| Pronoun types | 0.860 | 10 | 8.60 | 7.69 | 6.62 |
| Subordinate clause types | 0.588 | 7 | 4.12 | 7.69 | 4.52 |
| Clause finiteness | 0.742 | 7 | 5.19 | 7.69 | 5.71 |
| Number of matrix clauses in a paragraph | 0.741 | 5 | 3.71 | 7.69 | 5.71 |
| Function of a morpheme | 0.485 | 3 | 1.46 | 7.69 | 3.73 |
| Dictionary form of a word's root | 0.456 | 3 | 1.37 | 7.69 | 3.51 |
| Comma | 0.665 | 5 | 3.32 | 7.69 | 5.11 |
| Overall averages | | | 69.7 | $p=6.5*10^{-10}$ | 67.6 |

Green colors indicate topics in which the students perform better than the overall average, and red colors highlight topics in which the students perform below the overall average according to the original grading method. The average contributions are the products of the average scores and the corresponding weight factors. The average scores of the topics have been computed by dividing the total number of correct answers within the given topic by the number of students (68) and the number of questions within the given topic.

The very low p value indicates that the difference between the two methods is statistically significant even if not large. 13 students would receive more points according to the % method than according to the original grading method; however, only 1 would be given a higher grade – as mentioned above. The % method is clearly better suited for students who do roughly equally well in

all topics. Those who are good at clause constituents would be especially penalized unless they could compensate in the other topics. However, as can be seen from the distribution of points among the topics in Table 6, it is relatively seldom the case since the topic of clause constituents is below the overall average.

Another interesting result is that not one single student would gain more points from the methods that differentiate between questions according to their level of difficulty. In fact, everybody would gain fewer points from all the differentiating methods. This suggests that the students tend to answer the questions which are easier even when they are able to manage a difficult question. Nevertheless, it has been shown by simulations that it would be possible for a student to score up to grade 7 according to all the differentiating methods while still failing according to the original grading method provided they answer the more difficult questions instead of the less difficult ones. However, this opportunity would not have been exploited by any member of the sampled students.

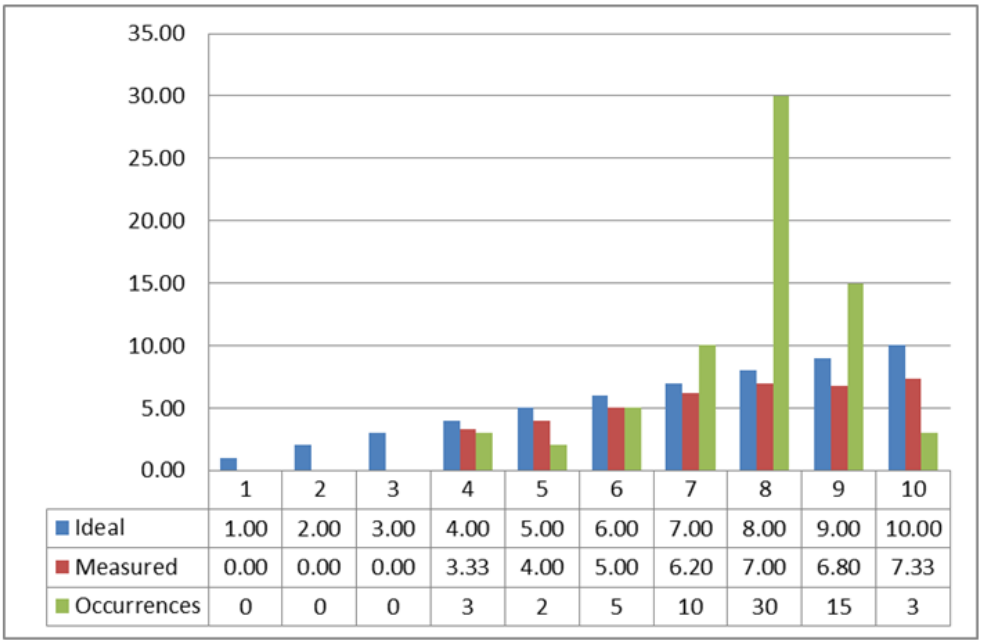
As for the differences between the differentiating methods, 9 students would receive more points – only marginally, though – according to the prop method than according to the ordinal method. The others (59) would receive fewer points as is reflected in the lower average. 16 students would receive more points according to the ordinal% and prop% methods than according to the ordinal and prop methods, respectively, but not the same 16 students.

In order to validate the differentiating methods, i.e. to ascertain that it is indeed reasonable to assign different levels of difficulty and therefore different numbers of points to the exam questions, the implicationality of the proposed difficulty levels has been assessed as well. The following figures show the implicational scales found for each topic.

'Ideal', repeated in each figure to ease the visual assessment of the scales, shows the ideal case of implicationality. 'Measured' shows the measured values for each question in the topics. 'Occurrences' shows the number of students who are at a particular level of difficulty. If the sum of occurrences is lower than the number of informants (68), it indicates that some students were not able to answer any of the questions in the particular topic correctly. There is no expectation as to the distribution of the occurrences. Therefore, it is not considered unusual, unexpected or undesirable if, for instance, a large number of students manage the question with the highest level of difficulty. It simply suggests that that question is not particularly difficult even though it is the most difficult one within its grammar topic.

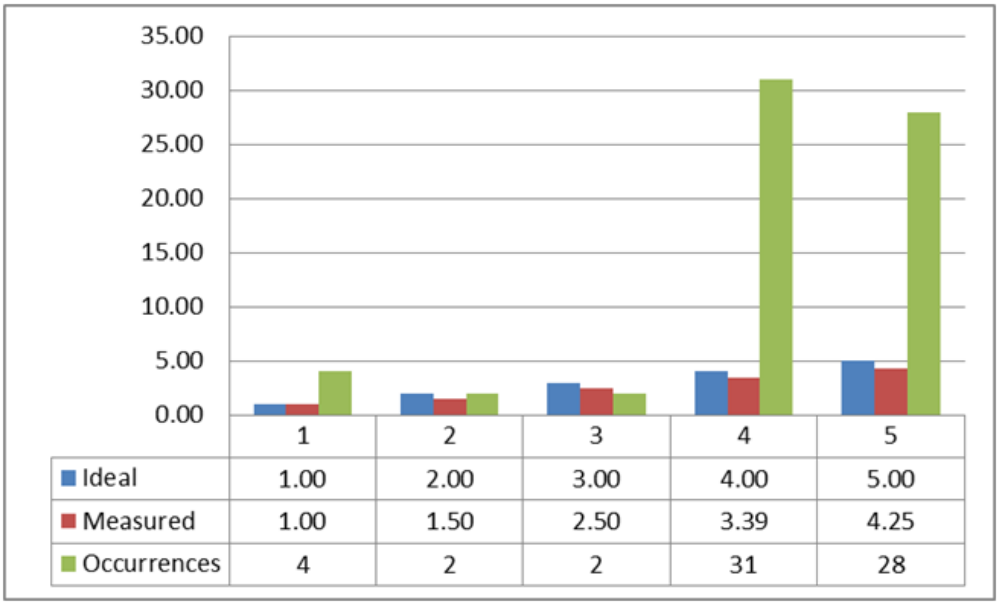
It must be noted that for the purpose of this paper, implicationality does not require that all the levels of difficulty be represented by at least one student. The scale is considered implicational if the measured values show a monotonously increasing tendency from any possible starting point, ignoring unattested levels of difficulty. The degree of implicationality of the scale, then, depends on how close the measured values come to the ideal values.

Figure 2: Parts of speech



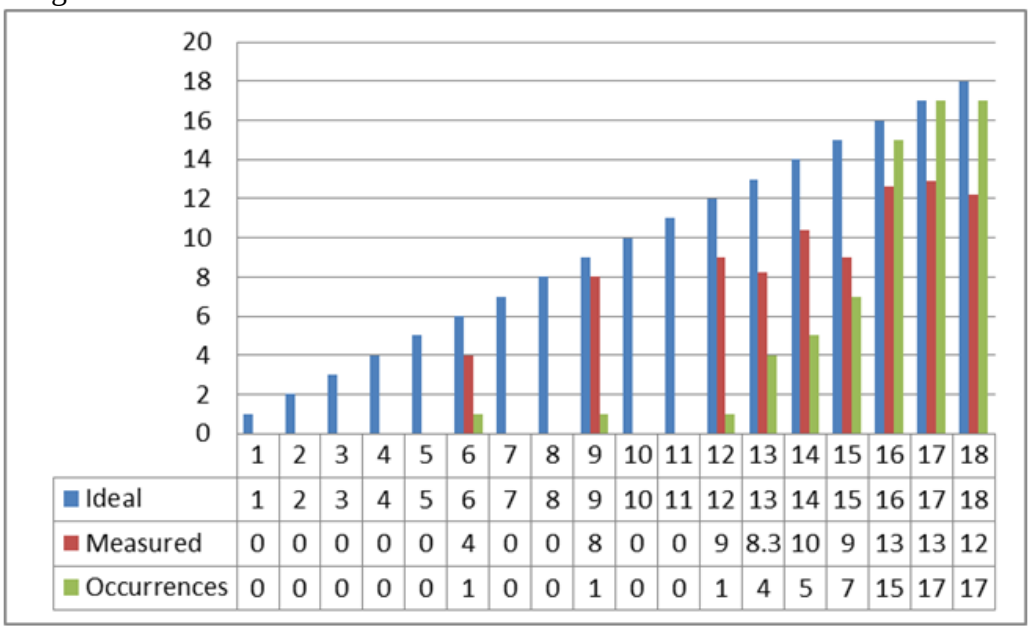

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Ideal | 1.00 | 2.00 | 3.00 | 4.00 | 5.00 | 6.00 | 7.00 | 8.00 | 9.00 | 10.00 |
| Measured | 0.00 | 0.00 | 0.00 | 3.33 | 4.00 | 5.00 | 6.20 | 7.00 | 6.80 | 7.33 |
| Occurrences | 0 | 0 | 0 | 3 | 2 | 5 | 10 | 30 | 15 | 3 |

Figure 3: Semantic relations



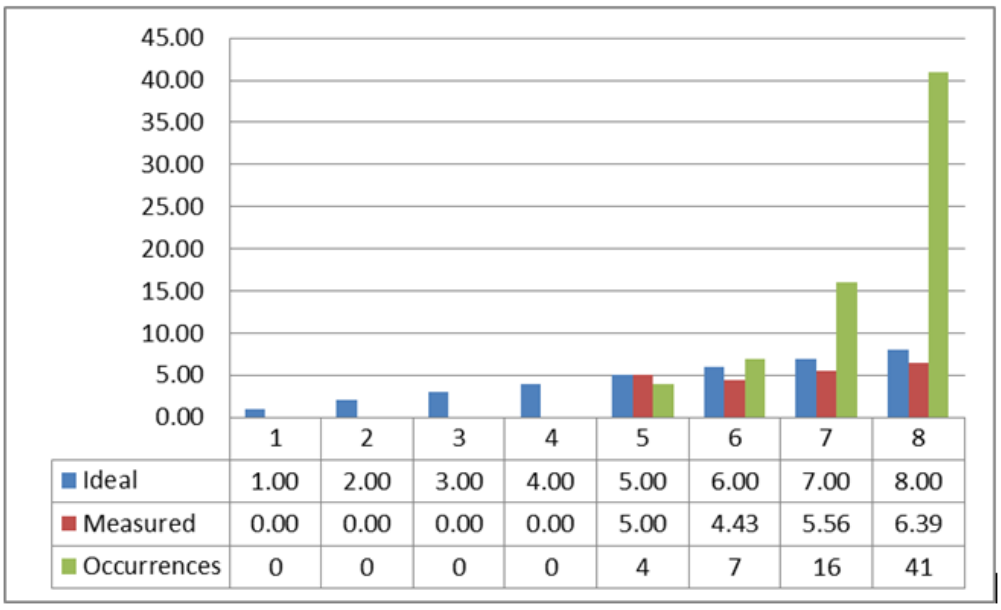

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Ideal | 1.00 | 2.00 | 3.00 | 4.00 | 5.00 |
| Measured | 1.00 | 1.50 | 2.50 | 3.39 | 4.25 |
| Occurrences | 4 | 2 | 2 | 31 | 28 |

Figure 4: Clause constituents



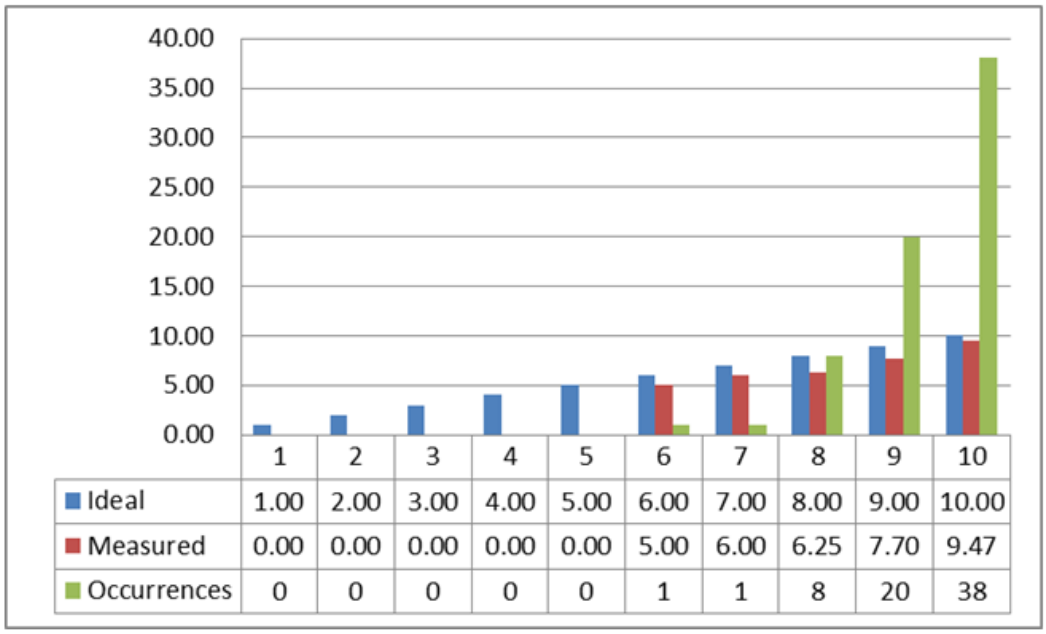

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ideal | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
| Measured | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 8 | 0 | 0 | 9 | 8.3 | 10 | 9 | 13 | 13 | 12 |
| Occurrences | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 4 | 5 | 7 | 15 | 17 | 17 |

Figure 5: Phrases vs subclauses



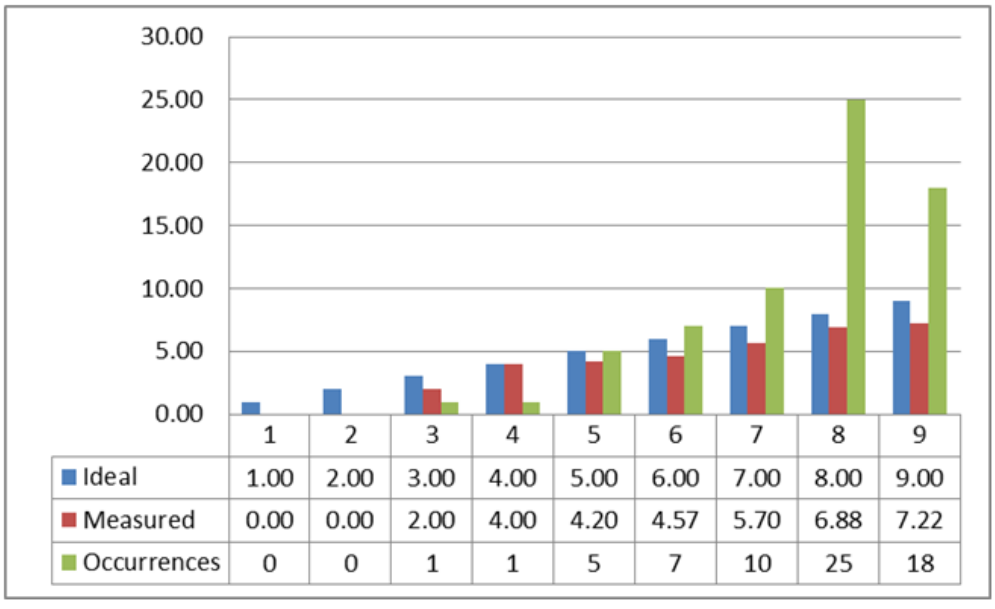

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| Ideal | 1.00 | 2.00 | 3.00 | 4.00 | 5.00 | 6.00 | 7.00 | 8.00 |
| Measured | 0.00 | 0.00 | 0.00 | 0.00 | 5.00 | 4.43 | 5.56 | 6.39 |
| Occurrences | 0 | 0 | 0 | 0 | 4 | 7 | 16 | 41 |

Figure 6: Phrase types



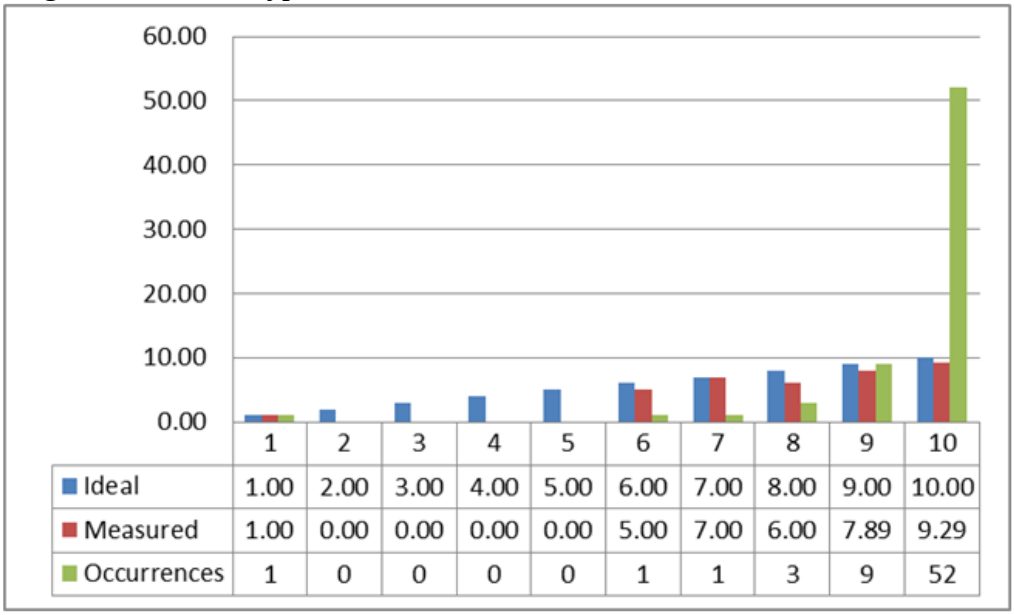

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Ideal | 1.00 | 2.00 | 3.00 | 4.00 | 5.00 | 6.00 | 7.00 | 8.00 | 9.00 | 10.00 |
| Measured | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 5.00 | 6.00 | 6.25 | 7.70 | 9.47 |
| Occurrences | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 8 | 20 | 38 |

Figure 7: Phrase constituents



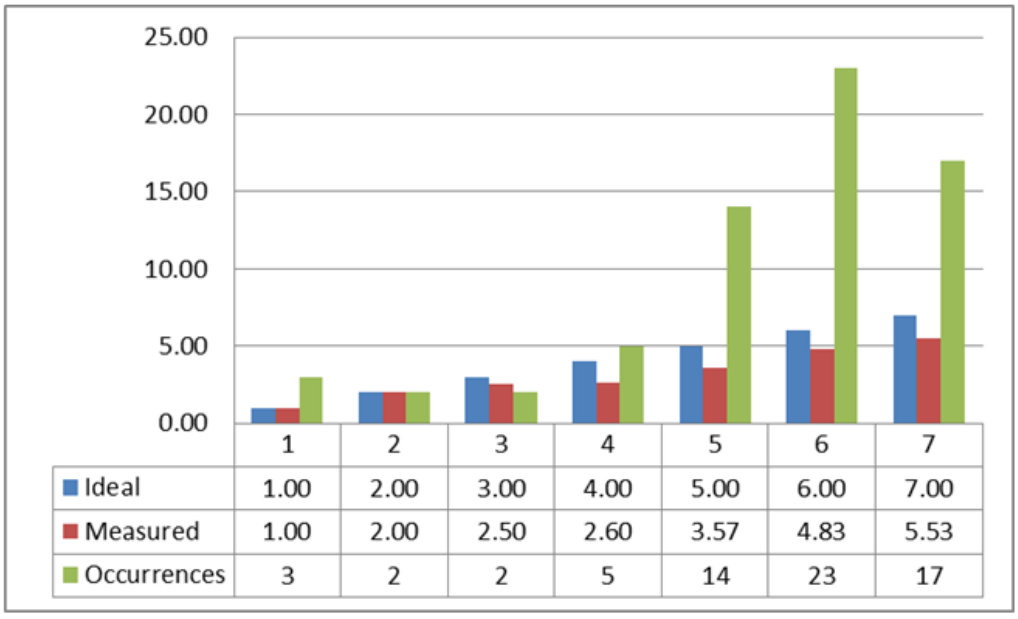

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| Ideal | 1.00 | 2.00 | 3.00 | 4.00 | 5.00 | 6.00 | 7.00 | 8.00 | 9.00 |
| Measured | 0.00 | 0.00 | 2.00 | 4.00 | 4.20 | 4.57 | 5.70 | 6.88 | 7.22 |
| Occurrences | 0 | 0 | 1 | 1 | 5 | 7 | 10 | 25 | 18 |

Figure 8: Pronoun types



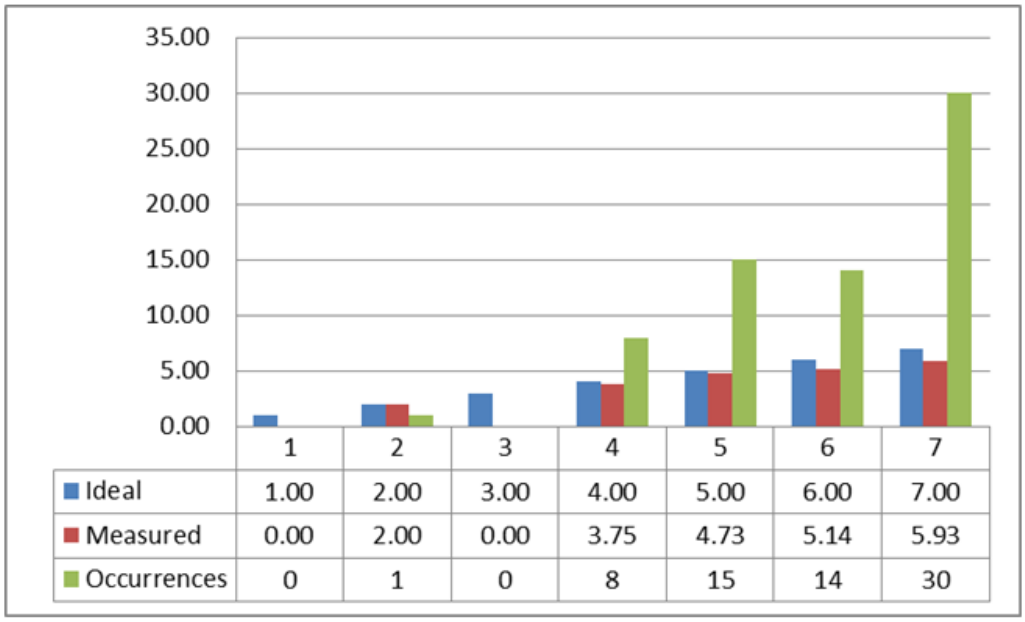

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Ideal | 1.00 | 2.00 | 3.00 | 4.00 | 5.00 | 6.00 | 7.00 | 8.00 | 9.00 | 10.00 |
| Measured | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 5.00 | 7.00 | 6.00 | 7.89 | 9.29 |
| Occurrences | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 3 | 9 | 52 |

Figure 9: Subclause types



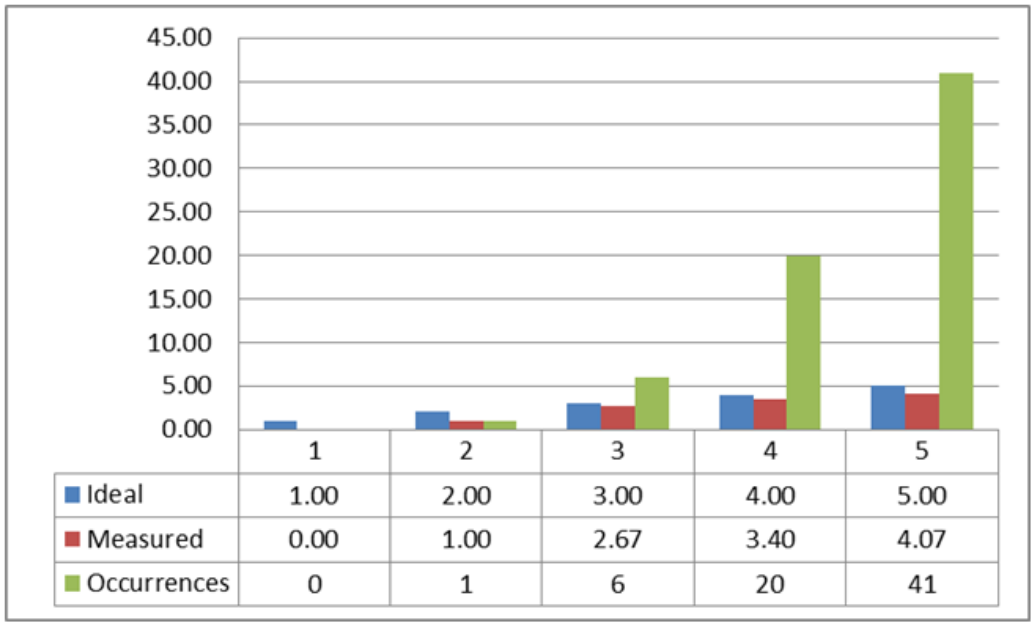

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Ideal | 1.00 | 2.00 | 3.00 | 4.00 | 5.00 | 6.00 | 7.00 |
| Measured | 1.00 | 2.00 | 2.50 | 2.60 | 3.57 | 4.83 | 5.53 |
| Occurrences | 3 | 2 | 2 | 5 | 14 | 23 | 17 |

Figure 10: Subclause finiteness



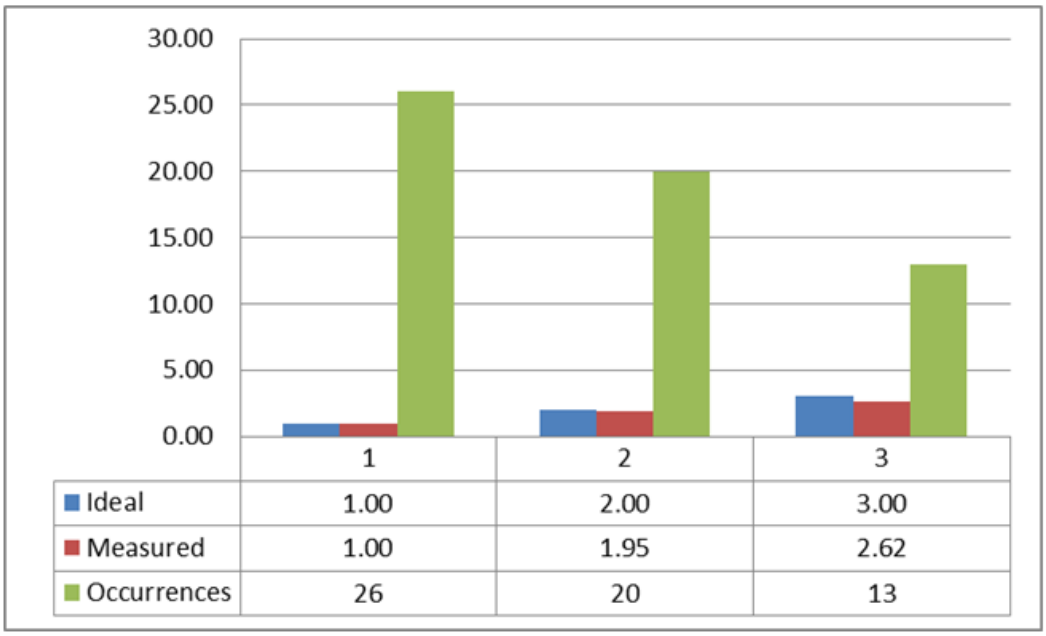

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Ideal | 1.00 | 2.00 | 3.00 | 4.00 | 5.00 | 6.00 | 7.00 |
| Measured | 0.00 | 2.00 | 0.00 | 3.75 | 4.73 | 5.14 | 5.93 |
| Occurrences | 0 | 1 | 0 | 8 | 15 | 14 | 30 |

Figure 11: Number of matrix clauses



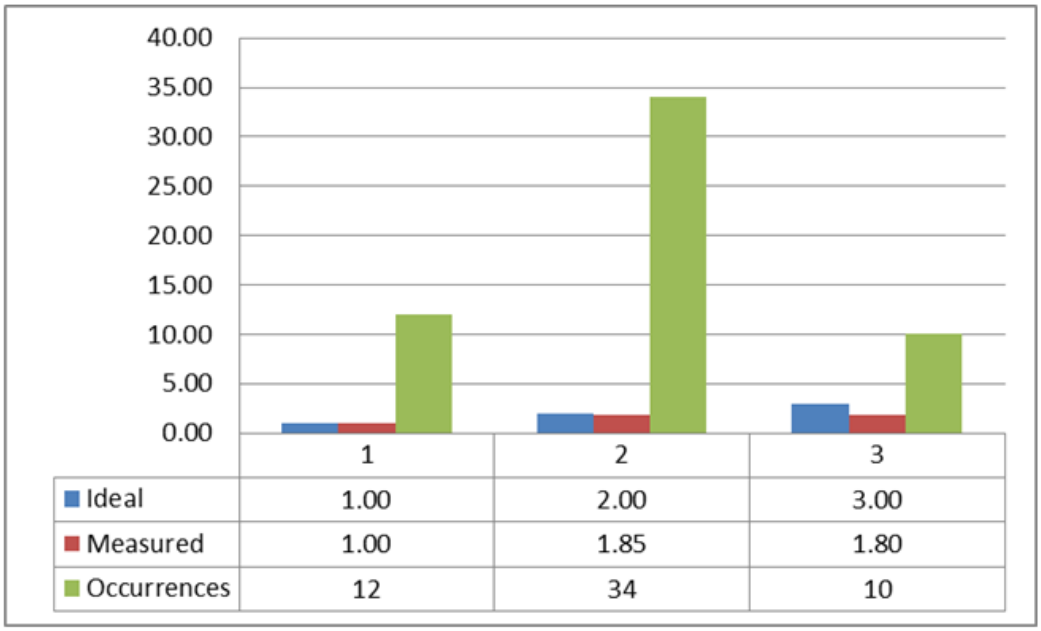

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Ideal | 1.00 | 2.00 | 3.00 | 4.00 | 5.00 |
| Measured | 0.00 | 1.00 | 2.67 | 3.40 | 4.07 |
| Occurrences | 0 | 1 | 6 | 20 | 41 |

Figure 12: Morpheme function



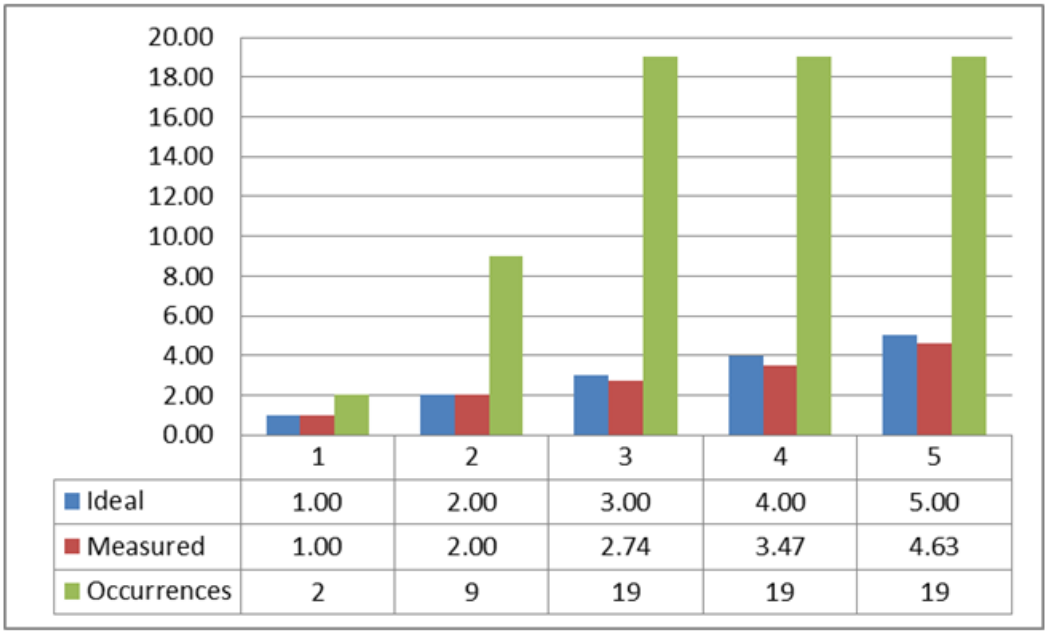

| | 1 | 2 | 3 |
|---|---|---|---|
| Ideal | 1.00 | 2.00 | 3.00 |
| Measured | 1.00 | 1.95 | 2.62 |
| Occurrences | 26 | 20 | 13 |

Figure 13: Word roots




| | 1 | 2 | 3 |
|---|---|---|---|
| Ideal | 1.00 | 2.00 | 3.00 |
| Measured | 1.00 | 1.85 | 1.80 |
| Occurrences | 12 | 34 | 10 |

Figure 14: Comma




| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Ideal | 1.00 | 2.00 | 3.00 | 4.00 | 5.00 |
| Measured | 1.00 | 2.00 | 2.74 | 3.47 | 4.63 |
| Occurrences | 2 | 9 | 19 | 19 | 19 |

As can be seen in the figures above and Table 7 below, which summarizes some parameters of the measured scales, the difficulty of most topics is approximately implicational. It justifies the attempt to differentiate between the exam questions based on the relative difficulty of the questions which is indicated by the numbers of students who answered the questions correctly.

'$P_{err}$', for proportion of errors, in Table 7 counts the number of incorrect answers to questions that should have been answered correctly by the students based on the highest level of difficulty they managed. This count is, then, divided by the number of questions in the given topic and by the total number of students. This value is essentially the aggregate difference between the ideal scale and the observed scale disregarding the gaps in the observed scale. Lower values are better, i.e. the observed scale is closer to the ideal one. Since these scales are not meant to be Guttman scales, this value is used instead of the coefficients of reproducibility and scalability (McIver & Carmines 1981).

'Avg. dist.' is the average of the differences between the blue and red bars in the figures, i.e. the average distance between the ideal pattern of responses and the observed pattern of responses to the individual questions, again disregarding the gaps in the observed scales. Lower values are better. 'St. dev.' is the standard population deviation of the aforementioned distances. Lower values are better because they indicate a more linear and uniform distribution of the observed responses. 'Inflx.' are the number of inflection points, where the linearity of the observed scale is broken, i.e. where a value on the ordinate is lower than the value immediately to its left (McMullen 2018). It should ideally be zero.

Table 7: Summary of scales of difficulty

| Topics | $P_{err}$ | Avg. dist. | St. dev. | Inflx. |
|---|---|---|---|---|
| Parts of speech | 12.9% | 1.33 | 0.716 | 1 |
| Semantic relations | 12.4% | 0.591 | 0.103 | 0 |
| Clause constituents | 24.9% | 3.74 | 1.56 | 3 |

| Topics | $P_{err}$ | Avg. dist. | St. dev. | Inflx. |
|---|---|---|---|---|
| Phrase vs. subordinate clause | 18.4% | 1.15 | 0.670 | 1 |
| Phrase types | 9.12% | 1.12 | 0.403 | 0 |
| Phrase constituents | 14.4% | 1.06 | 0.522 | 0 |
| Pronoun types | 7.94% | 0.965 | 0.646 | 1 |
| Subordinate clause types | 16.8% | 1.00 | 0.555 | 0 |
| Clause finiteness | 10.5% | 0.489 | 0.404 | 0 |
| Number of matrix clauses in a paragraph | 15.6% | 0.715 | 0.267 | 0 |
| Function of a morpheme | 2.90% | 0.217 | 0.167 | 0 |
| Dictionary form of a word's root | 8.33% | 0.674 | 0.526 | 1 |
| Comma | 6.47% | 0.289 | 0.192 | 0 |

The topic of clause constituents is rather messy. It has several inflection points, and the distances between the measured and ideal values are also large, much larger than the corresponding values of the other topics. Whether this extraordinary behavior of clause constituents is an artefact of the particular questions asked in the exam investigated here, of the sample group of informants, or is intrinsic to this grammatical category (having perhaps a multidimensional nature) cannot be ascertained without calculating the implicationality of other questions on clause constituents and testing other informant groups.

However, some signs suggest that there is something "fishy" about clause constituents. Madsen (2017) has found that clause constituents are consistently among the topics that students find most challenging despite the fact that it is among the few grammar topics that are already taught in primary school. Most of the grammar topics examined are new to the students when they enter the university, yet these behave fairly "normally". Furthermore, mastering clause constituents seems to be one of the major factors in learning theoretical grammar (Madsen 2015). Looking at clause constituents more closely will therefore make a good theme for a follow-up paper.

## 6. Conclusion

It has emerged that none of the proposed grading methods would be in favor of the students in terms of grades, as all the new methods would yield the same or lower grades than the method used thus far, with one single exception. The reason for this is that the students have a strong tendency to answer the easier questions instead of the more difficult ones. It is, of course, not surprising as a general tendency, especially in view of the relatively low average score of 69.7. However, it was unanticipated that so few would benefit from the alternative methods. Whether this fact speaks for or against implementing one of the alternative grading methods is a political question, and the methods should also be tested on further samples.

In any case, the paper revealed both positive and negative aspects of the exam investigated, which are worth considering both for further research and teaching/examining of grammar. On the positive side, most topics are probed by questions which form an implicational scale of difficulty even though it has never been intentional. It justifies the use of a grading system, such as the ones tested in this paper, which differentiates between questions based on their degree of difficulty. Furthermore, it is a good starting point for refining future exams if it is desired that the exam questions be (more) implicational with regard to difficulty. On the negative side, there is a discernible bias in

the exam towards certain topics, and it ought to be a source for reflection whether this situation should be upheld on some principled ground or abolished in future exams.

On the surprising side, the topic of clause constituents has been found rather fuzzy. It begs for further investigation not only for the sake of the development of exams and their grading, but also for the research in grammar acquisition and the teaching of grammar. As for the exams, the fuzziness of clause constituents is problematic since this topic is the one that is weighted most in the current way of examining grammar. This can create unwanted and uncontrollable bias. As for the learning and teaching of grammar, clause constituents may pose yet unrealized challenges, which may not only be pertinent to this part of grammar.

**References**

Aitken, Nola (2016). 'Grading and Reporting Student Learning' in Scott, Shelleyann, Donald E. Scott & Charles F. Webber (eds.) *Assessment in Education*. Heidelberg: Springer International Publishing. 231-260.

Bettoni, Camilla & Bruno Di Biase (2015). *Grammatical development in second languages: Exploring the boundaries of Processability Theory*. Amsterdam: The European Second Language Association.

Bovey, Rob, Dennis Wallentin, Stephen Bullen & John Green (2009). *Professional Excel Development: The Definitive Guide to Developing Applications Using Microsoft Excel, VBA, and .NET*. Upper Saddle River, NJ: Pearson Education. Kindle Edition.

Brookhart, Susan M. (2011). *Grading and Reporting: Practices that Support Student Achievement*. Bloomington, IN: Solution Tree Press.

Carlberg, Conrad (2014). *Statistical Analysis: Microsoft Excel 2013*. Indianapolis: Que Publishing. Kindle Edition.

DeVellis, Robert F. (2011). Scale Development: Theory and Applications (Applied Social Research Methods). London: SAGE Publications. Kindle Edition.

Donato, Richard (1994). 'Collective scaffolding in second language learning'. In Lantolf, James P. & Gabriela Appel (eds.), *Vygotskyan Approaches in second language research*. New York: Ablex Publisher. 33-56.

Dörnyei, Zoltán (2014). Questionnaires in Second Language Research: Construction, Administration, and Processing. Abingdon: Taylor and Francis. Kindle edition.

Hatch, Evelyn & Hossein Farhady (1982). *Research Design and Statistics for Applied Linguistics*. Rowley, MA: Newbury House Publishers.

Hjulmand, Lise-Lotte & Helge Schwarz (2008). *A Concise Contrastive Grammar of English for Danish Students*. Frederiksberg: Samfundslitteratur.

Lehmann, Christian (2018). *Gewaltenteilung in der Bildung*. http://privat.christianlehmann.eu/ (accessed 02.13.2019).

Madsen, Richard (2015). 'A statistical model of learning descriptive grammar'. In Anna Bondaruk, Anna Bloch-Rozmej, Wojciech Malec, Ewelina Mokrosz and Sławomir Zdziebko (eds.), *Young Minds vs. Old Questions in Linguistics: Proceedings of the Fourth Central European Conference in Linguistics for Postgraduate Students*. Lublin: The Institute of East-Central Europe and the John Paul II Catholic University of Lublin. 122-138.

Madsen, Richard Skultety (2017). What is wrong with grammar? Danish university students' difficulties with the acquisition of written English and theoretical grammar. Aalborg: Aalborg Universitets Forlag.

Madsen, Richard Skultety (ms). 'Learning curve. Can the exam scores of the grammar exam be predicted?'.

McIver, John & Edward Carmines (1981). *Unidimensional scaling. Quantitative Applications in the*

*Social Sciences 24*. Newbury Park, CA: Sage Publications.
McMullen, Chris (2018). Essential Calculus Skills Practice Workbook with Full Solutions. Zishka Publishing. Kindle Edition.
Ministry of Education (2017). *7-point grading scale*. http://eng.uvm.dk/general-overview/7-point-grading-scale (accessed 02.13.2019).
Ministry of Education (2018). *The Danish education system*. https://ufm.dk/en/education/the-danish-education-system (accessed 14.02.2019).
Ministry of Education (2019). *7-trins-skalaen*. https://uvm.dk/uddannelsessystemet/7-trins-skalaen/anvendelse-af-7-trins-skalaen (accessed 02.12.2019).
Vygotsky, Lev (1978). Mind in society: The development of higher psychological processes. Cambridge, MA: Harvard University Press.