Globe: A Journal of Language, Culture and Communication, 11: 43-58 (2020)

The learning curve.
Can the results of the grammar exam be predicted?

Richard S. Madsen, Aalborg University

Abstract: This paper exploits the potentials of inferential statistics in its quest to answer two related questions, whether the exercises performed by students during their course in theoretical grammar really prepare them for the exam, and whether the students' exam results can be predicted from their achievements in said exercises. This study is in the context of English Business Communication at Aalborg University, Denmark. Several statistical methods, various forms of regression analysis, are pursued in order to discover which one – if any – of them is best suited to make predictions as to future exam results. It is found that the exercises investigated do indeed contribute significantly to the students' learning process, and that the exam results are predictable within a reasonable margin of error from the results of the exercises. Somewhat surprisingly, the simplest forms of linear regression and 1st degree polynomial regression are found to be the best predictors of the exam results, not any of the elaborate methods also tested. As a side effect, the study also reveals that the students' level of knowledge of theoretical grammar prior to their entering the university has no appreciable influence on their exam results.

Keywords: Learning of grammar, inferential statistics, language acquisition.

1. Introduction

It is a common and reasonable expectation that the exercises that the students in a course are required to do prepare the students for their exam. It is no less the case with a course on theoretical grammar. The purpose of this paper is to investigate to what extent this expectation is fulfilled in the case of the English Grammar course which first-semester students of English Business Communication at Aalborg University have to absolve. A related issue that is investigated is how predictable the exam results are from the performance of students during the course as documented through the exercises they do during the course. The purpose of investigating the predictability of the result of the grammar exam is to give the course teacher a tool to detect students who may be at risk of failing the exam.

The course investigated in this paper is a course on theoretical or descriptive grammar, in which the students have to learn concepts of grammar such as clause and phrase constituents, affixes and parts of speech, etc. This course has thus its focus on theoretical concepts applicable to English grammar. A separate course called Production of Written Texts focuses on the use of English, in which the students learn to employ grammar properly in practice. These two courses are of course tightly knit; so much so that they have a common portfolio exam at the end of the second semester. Also, the study regulation stipulates that learning theoretical grammar will improve the students' mastery of English and their grammatical precision in using English (Study Board of Language and International Business Communication 2017). However, this paper concerns itself only with how well the students are prepared for their exam in theoretical grammar. It does not address the issue of how much of that knowledge is eventually converted into a practical command of English. For that question, see Madsen (2014).

2. Inferential statistics

This paper relies on the theory of inferential statistics to answer the two related research questions "how well the home assignments that the students do during the course prepare them for the course-final exam" and "how predictable the result of the grammar exam is". The study uses correlation analysis to answer the first question, and multiple linear regression and polynomial regression to answer the second question (Urdan 2012; Hartshorn 2017). The pursuit of the second question has been inspired by Elbro and Scarborough (2003). This section explains why these technics have been

selected, Section Data describes the data used in detail, and Section Method explains how the data were manipulated in the analysis. A good understanding of the nature of the data is necessary for the understanding of the data manipulation employed.

2.1 Correlation analysis

The correlation analysis is used to determine the level of correspondence between the results of the home assignments during the course and the exam results, with a high level of correlation suggesting that the assignments prepare the students well for the exam. The hypothesis is the following: If the students become prepared for the exam during the grammar course, then the level of correlation is low at the beginning of the course, is increasing towards the end of the course and reaches a high level for the exam. Low correlation at the beginning corresponds to the relatively low level of knowledge the students enter the university with. Increasing correlation indicates the rising level of knowledge of the students, and a high final level of correlation indicates that the course has prepared the students well for the exam. Should the correlation show a decreasing tendency during the course, it would indicate that the course is detrimental for the students' development. Low final correlation and/or non-increasing level of correlation would suggest that the students do not gain much from doing the exercises with respect to the exam.

Note that the level of correlation and its tendency during the course are not indicative of the students' level of knowledge, only of how well the assignments' results tally with the exam's results. The level of knowledge at any stage is only indicated by the results of the exam and the home assignments, respectively. Thus, in principle, the students can attain a high level of knowledge even without a significant level of correlation between the home assignments and the exam, and, similarly, a high level of correlation may be accompanied by a low level of knowledge.

2.2 Regression analysis

The purpose of the regression analyses is to devise an equation, a mathematical formula, that can be used to compute the expected score at the exam of a future student. This formula takes a student's scores in the home assignments and the course-initial test as input and outputs the score that the student is likely to achieve at the exam. The result of this calculation can then reveal whether the student is in danger of failing, in which case evasive actions can be taken in time, before attempting the exam. Since this study attempts to predict only one value, the aggregate score of the exam, regression analysis suffices. Were the purpose to predict several values, for example subscores of the exam (cf. Section Data), multivariate analysis would be called for (Hair et al. 2010).

Of the several different ways of building such an equation, two kinds of regression analyses have been performed, multiple linear and polynomial (Boundless 2014). Since they both have advantages and disadvantages, one of the purposes of this study is to evaluate which one merits future use or further study.

The multiple linear regression analysis takes a sample group of students' scores in the home assignments and the course-initial test as the independent variables and the scores in the exam of the same group of students as the dependent variable and builds an equation that relates the dependent variable to the independent ones. This equation can then be used to predict the expected score at the exam of a student who is not a member of the sample group.

This is the cornerstone of this kind of analysis; the equation is deployed on students whose data were not used in the development of the equation. Hence, the success of the equation in predicting a student's exam scores depends heavily on how similar the student behaves compared to the sample group. The more dissimilar they are, the more imprecise the prediction becomes. In fact, the reliability of the equation also depends on the homogeneity of the sample group. The more heterogeneous the sample is, the less reliable the equation becomes. This dependence on similarity is a disadvantage of this method. However, it can be countered by the possibility that the input data can

be readily modified in a multitude of ways in order to fine-tune the equation. In fact, the multiple linear regression analysis has been performed on the same set of initial data modified in six different ways in order to find the data manipulation best suited.

The polynomial regression analysis has the advantage that it does not require a sample group of students. Hence, the equation is developed on the data of the very same student whose future exam score it is supposed to predict. In this way, a unique equation is developed for every student whose exam score is to be predicted – as opposed to the one-size-fits-all approach of the multiple linear regression analysis. The disadvantage of this method is that it depends on the consistent performance of the student for whom the equation is developed. If he or she does not perform in a mathematically detectable consistent pattern during the course, the equation gained is likely to mispredict the exam result of the student.

3. Data

As mentioned in Section Inferential statistics, the data for the analyses are the students' scores from the course-final exam, from the three compulsory home assignments during the course and from an extracurricular course-initial test that was used to gauge the students' knowledge of grammar upon entering the university.

The exam the students have to take in grammar at the end of their first semester consists of 95 questions on 12 topics within theoretical grammar[1]. An additional five questions concern the use of comma in certain sentences; however, these five questions were taken out of the dataset. 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.

The three compulsory home assignments during the course also contain 100 questions each and have to be done within one week with intervals of two weeks. The students are allowed to use any means of aid except human help. Possible plagiarism is actively checked and punished when proven beyond reasonable doubt. The course-initial test, containing 24 questions, has to be completed within 20 minutes without any aid. All the questions in the home assignments and the course-initial test are from previous exams with occasional minor modifications.

Table 1 gives an overview of the topics of grammar discussed during the first semester and how they are weighted in the course-initial test, the home assignments and the exam[2].

Table 1: Overview of the grammar topics taught and examined

Number of questions in Topics	Course-initial test	Home assignment 1	Home assignment 2	Home assignment 3	Exam
A. Parts of speech	9	40		10	10
B. Semantic relations			8	3	5
D. Clause constituents	15	10	25	20	18
E. Phrase vs. subordinate clause			15	10	8
F. Phrase types			15	10	10
G. Phrase constituents			12	10	9

[1] This description applies only to the grammar exam in English. The German and Spanish groups of International Business Communication in Aalborg follow a different approach.

[2] The reason for the gaps in the alphabetical codes of the topics is that the topics which are coded with the letters C, M, N, O, U, V, W, X and Y had either been deprecated before this investigation was performed or are only discussed and evaluated in the second semester and are thus out of the scope of this paper. L and Z were removed from the analysis (see Section 4 for explanation).

Number of questions in Topics	Course-initial test	Home assignment 1	Home assignment 2	Home assignment 3	Exam
H. Pronoun types		10		10	10
I. Subordinate clause types				7	7
J. Verb finiteness		15	7	7	7
K. Number of matrix clauses in a paragraph				5	5
L. Comma with relative clauses				(5)	
P. Number of affixes in a word		10			
Q. Part of speech of a word's root		10			
R. Function of a morpheme		5	8		3
S. Number of constituents in a phrase			10		
T. Dictionary form of a word's root				3	3
Z. Comma					(5)

As can be seen in Table 1, only home assignment 3 is almost like the exam; the other assignments and especially the course-initial test deviate from the exam's structure considerably. However, all the assignments and the test consist of questions from exams in previous years. The reason for the structural variation is many-fold and explained below.

The course-initial test contains questions only on parts of speech (A) and clause constituents (D) because these are the only topics that can be expected to be known by all students emerging from secondary education in Denmark. Some students might have been introduced to more sophisticated grammar in high school; however, since such students make up a small minority, it would not have made sense to waste precious time on questions that most students would not have had a chance to answer.

Since most of the grammar topics are hence completely new to the students, and since according to the study regulation, all the home assignments must be submitted during the course, it is not possible to train all the topics in all the home assignments. Some topics (for example I) must wait until the end of the course to be discussed and can thus be included only in the final home assignment.

The reason for the high proportion of questions in home assignment 1 on parts of speech (A & Q) is partly that it is one of the topics that students are already familiar with and that this topic was found to be one of the most essential topics in theoretical grammar (Madsen 2015). A similar argument applies to the large number of questions in home assignment 2 on clause constituents (D). The reason for the relatively many questions on finiteness (J) in home assignment 1 is that this topic had been believed to be particularly difficult for the students. However, a later examination found it to be a topic in which the students consistently performed above the overall average (Madsen 2017). Despite their not appearing in further home assignments or in the exam, topics *P, Q* and *S* were used to give the students another perspective on morphological and phrase analysis, which are two realms of grammar completely new to the students. Apart from these pedagogical considerations, it was attempted to allot the various topics roughly equal representation in the home assignments and the

exam.

Table 2 provides some examples of the questions posed within the different topics.

Table 2: Examples of questions in the exam, home assignments and course-initial test

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

- If you want to drink a healthy alcoholic beverage, cider is a very good choice.

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

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

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

- True cider is made from fermented apple juice.

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

- The wide availability of apples makes it easy to produce cider almost anywhere.

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

- the alcoholic content of cider

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

- Whoever invented cider was a genius.

***I*&*J*. Determine the type and finiteness of the underlined subclauses.**

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

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

- Unhealthily

With the exception of question type *T*, the students have to select the correct answer from fixed sets of valid answers. For instance, in the case of *D*, the set of valid answers is the set of clause constituents, containing nine elements in this grammar course (Hjulmand & Schwartz 2012). Should a student give a true but invalid response, say calling *from fermented apple juice* a preposition phrase instead of 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. It does happen, however, that some students provide invalid responses. In question type *T*, there is no fixed set of valid responses, and the students are not given any hints as to what the root might be except for the word itself.

Each correct and valid response yields one point for the student. Incorrect and non-existent responses yield zero points. The students have to answer 60% of the questions correctly in order to pass the exam. The course-initial test and the home assignments are not graded. Nevertheless, the students have to make all the home assignments to be allowed to attempt the exam although the results of the home assignments do not matter.

4. Method

This section gives a detailed description of the setup of the regression analyses. The correlation analysis was done in a straightforward manner, not requiring a lengthy elaboration, and is summarized in the last subsection. All the calculations were performed in MS Excel (Bovey et al. 2009; Falls

2011; Carlberg 2014; Harmon 2014).

As mentioned in Section Inferential statistics, the students' scores at the course-final exam serve as the dependent variable, and their scores in the course-initial test and in the three home assignments serve as the independent variables in the multiple linear regression analysis. The scores of freshmen in 2014 (60 students) and 2015 (79 students) were used in this study.

The years 2014 and 2015 were used because the students in these years were given exactly the same course-initial test and home assignments. The exam questions were of course different token-wise in the two years, but the same type-wise. This setup ascertains the best possible way to test the equations' predictive power. However, the equations gained in this way require for maximum dependability that they be used on the scores of the same home assignments in the future that they have been developed on.

The five questions on comma in the exam were disregarded in the analysis because, due to a human error, they were rendered unusable in the 2015 exam. Thus, that exam featured only 95 questions. For the sake of comparability, the results of the 2014 exam were adjusted so as to ignore the questions on comma. Similarly, question type *L* was taken out from the results of the third home assignments in both years. This has to be taken into account when attempting to predict a student's exam result. The equations developed will not make a prediction with regard to questions on comma use. However, since this paper focuses on the learning of theoretical grammar, it is arguably not a major loss that comma use has been ignored.

4.1 Multiple linear regression

The multiple linear regression analysis, i.e. the development of the equation to predict the exam score, was done separately for the two years, and then the equations were checked against each other's samples in order to test their predictive power. In other words, the equation developed for the students in 2014 was checked against the exam scores of the students in 2015, and the equation developed for the students in 2015 was checked against the exam scores of the students in 2014.

As mentioned in Section Inferential statistics, different manipulations of the raw data were performed before doing the regression analysis. Thus, two times four pairs of predictive equations were developed, four pairs for either year because four different sets of input data were used. One member of each pair takes into account the course-initial test, and the other member of the pair does not. The reason for this was that the course-initial test was extracurricular, so it cannot be expected that future courses – possibly taught by other teachers – will employ it. On the other hand, the three home assignments were compulsory and can be expected to remain so. Having one equation that considers the course-initial test and another equation that does not also makes it possible to estimate how much the students' initial level of knowledge means for their achievement in the grammar course.

In the following, the four different input data sets are described. The simplest of the multiple linear regression calculations is based on the aggregated scores of the home assignments and the course-initial test, *i.e.* on the number of correct answers in each. Thus, these calculations take three or four factors into account (without and with the course-initial test, respectively). In the other three types of multiple linear regression calculation, the scores of the individual topics are taken separately as factors. Hence, these calculations operate on 27 and 25 factors depending on whether the course-initial test is or is not taken into consideration.

The reason for trying this more complex approach is that the aggregate scores do not reveal anything about the composition of the students' knowledge. Two students can have the same aggregate score, but they have likely answered different questions correctly, and their knowledge may thus have different compositions. Different knowledge compositions may have different bearings on the exam scores.

Table 3 summarizes the datasets for the multiple linear regression analysis.

Table 3: Datasets used in the study

aggregate scores of home assignments and course-initial test	scores of the individual topics within the home assignments and course-initial test		
	nominal scores	differentiated scores	
		ordinal	proportional

The simplest one of the three more detailed calculations (called nominal scores) takes the scores for the individual grammar topics at their face value, i.e. simply the number of correct answers. The other two approaches attempt to differentiate between the responses depending on the difficulty of the questions. The idea behind this is that the questions within the same topic are inevitably at different levels of difficulty, be it accidentally or purposefully. Hence, even if two students have answered the same number of questions correctly, they may have knowledge of different levels or of different composition within the grammar topic concerned. This difference may in turn have a bearing on the final score at the exam.

The level of difficulty of the questions is estimated from the number of correct responses given to the individual questions by the members of the sample group (Hatch and Farhady 1982). Thus, the measure of difficulty is dependent on the composition of the sample group on whose scores the equation is being built and may differ from group to group. However, since the difficulty of questions can be validly measured only by the performance of those who have actually answered the questions, there is no reliable independent external measure of difficulty. The only way to mitigate the effect of the possible difference between sample groups is to measure the questions' difficulty on the performance of one very large group of informants and use this measurement in all subsequent calculations. However, since this study is the first of its kind – at least for the students of English Business Communication – this pre-assessment of question difficulty was not possible.

Another beneficial effect of this differentiation is that the granularity of the scores increases, i.e. the same number of questions can differentiate between more students. Without the differentiation, the undifferentiated aggregate score of say five questions can only differentiate between six students because there are only six different scores: 0, 1, 2, 3, 4, 5. With a differentiating algorithm, this number can be increased to 20. The algorithm is explained in detail in Madsen 2019 and therefore only summarized here briefly.

Two methods of differentiation are used in this study as well. In one of them, an integer value from 1 through the number of questions within the topic (n) is assigned to each question depending on the detected level of difficulty. 1 indicates the lowest level of difficulty, and n the highest, the level of difficulty being inversely proportional to the number of informants having answered that question correctly. If two questions appear to have the same level of difficulty, then they are assigned the same value. This method is referred to as 'ordinal' in Table 3.

The other method of differentiating is called 'proportional' in Table 3 and works in the same way except for the fact that the level of difficulty is not expressed by integer, but by rational numbers. In this way, 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 rationale is that the difficulty of the questions is not necessarily equally distributed. For instance, the second easiest question may be as many as five times more difficult than the easiest question while the third easiest question only slightly more difficult than the second easiest one. Also

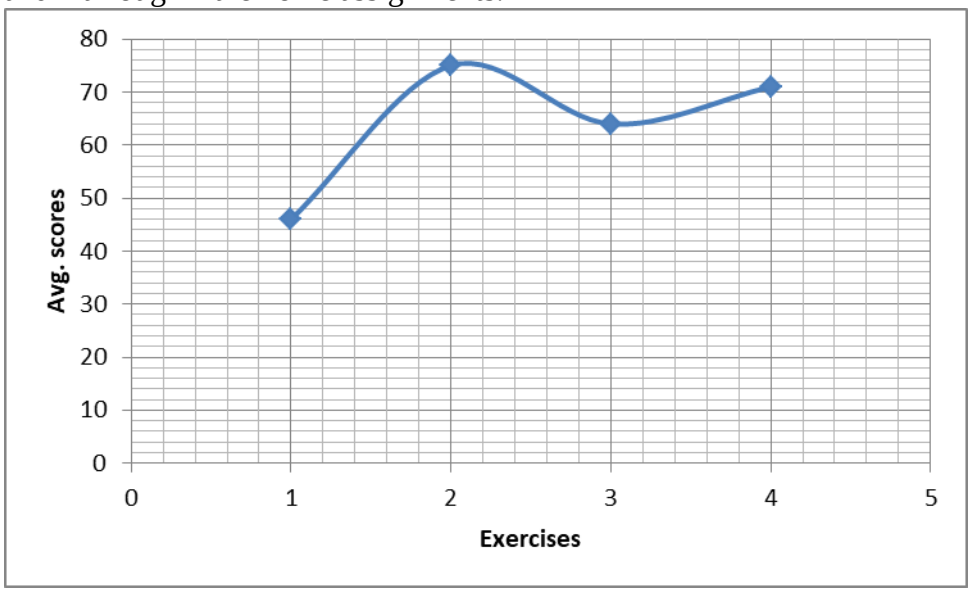
in the proportional method, the easiest question is assigned the value of 1, and the most difficult question the value *n*. All the other questions are assigned values between 1 and *n* in proportion to their measured difficulty.

For the sake of comparability with the aggregate and nominal scores (Table 3), the differentiated scores are scaled to the same ranges as those of the nominal scores. That is, if a topic is probed by 10 questions, then the maximum number of points attainable is 10 regardless of the calculation method. The differentiated scores just have a higher resolution of the same range than the nominal scores.

4.2 Polynomial regression

The idea of attempting a polynomial regression analysis comes from the observation (Figure 1) that when plotting in a coordinate system the sample means for the course-initial test and the home assignments in chronological order, the points resemble the curve of a 3rd degree polynomial (the curve is continuous for illustrative purposes only). Hence, this idea takes the title of this paper literally and attempts to fit a curve onto the data points. The equation that describes the curve can then be used to predict the score at the exam, which is the next data point in the curve.

Figure 1: 3rd degree polynomial of average scores. 1 on the *x* axis represents the course-initial test, and 2 through 4 the home assignments.



Admittedly, this is a rather dubious approach. Even though four points technically define a 3rd degree polynomial, it is not at all clear what the independent variable on the *x*-axis represents. The best bet is probably time; however, even so, it is not given what values it should assume. If the course-initial test and the home assignments are assigned the *x*-values of 1 through 4, should the exam be represented by 5? It was in the test calculations because given the nature of the cubic polynomial, any higher value would likely overshoot the target with a very large margin. However, it was an arbitrary choice, which in fact does not resonate well with the assumption that the independent variable should be time. For assuming the values of 1 through 5 suggests that the measurement points are equidistant in time. It does apply to the home assignments because they were spaced two weeks from the course-initial test and each other. However, the exam took place about one month after the last home assignment, which suggests that the *x*-value for the exam ought to be 6. A similar doubt concerns the dependent variable on the *y*-axis. Its values are seemingly given – the above mentioned scores; however, these values derive from different sources and only vaguely measure the same thing, the

student's level of knowledge.

Despite all these grave reservations, a 3rd degree polynomial was considered worth pursuing, possibly even as a better choice than 1st degree polynomial, *i.e.* a linear approximation. The reason why a 1st degree polynomial was considered possibly inferior is that most students showed a generally increasing tendency from the course-initial test to the third home assignment, and a linear approximation would necessarily carry this tendency further. However, not all students' exam results warrant this assumption. Therefore, it was expected that a 1st degree polynomial would have a strong tendency to overestimate the exam result, which is clearly undesirable.

In any case, both 1st and 3rd degree polynomials were tried, and both in two different ways. In one of the approaches, the starting point (the first data point) for the polynomial was the score of the course-initial test. In the other approach, the starting point was set at the origin, in a way modelling the absolute starting point of the students having zero knowledge of grammar. As opposed to the multiple linear regressions, the 3rd degree polynomials do not have different versions according to whether the course-initial test is included or not. Since a 3rd degree polynomial needs at least four data points, it would make no sense to disregard the course-initial test. Adding the origin as an extra starting point is only an artificial data point.

The polynomials' average performance was evaluated against the combined sample of the two years. Since each informant individually serves as the basis for a set of four polynomials, there is no point in comparing the years against each other. However, a larger sample can give a more reliable picture of the average performance of the polynomials.

4.3 Correlation analysis

The aggregate exam scores of the students were correlated with the aggregate scores in the home assignments and the course-initial test, one pair of datasets at a time. The two years were combined in the correlation analysis to increase the sample size. The significance of the correlations was calculated by using the two-tailed pairwise *t*-test. The correlation analysis is augmented by a calculation of the difference between the average of the exam scores and the averages of the exercise scores.

5. Analysis

The first two sections report the results of the multiple linear regression analysis and the results of the polynomial regression analysis, respectively. These results are evaluated according to the following metrics. The results of the correlation analysis are presented separately.

- $\bar{x}_E$ = mean of the expected scores. It is the arithmetic mean of the exam scores that are predicted by the equation developed by the regression analysis. It should be as close to the mean of the observed scores as possible.
- r_{EO} = correlation coefficient between the observed (or target) scores and the expected (or calculated) scores. It goes from -1 through +1. The closer it is to +1, the better it is from the perspective of this study. Values closer to 0 indicate weaker correlation. Negative values would indicate inverse correlation between the observed and expected scores, which would be catastrophic for the purpose of this study.
- p = probability of equality as calculated by the two-tailed pairwise *t*-test. It can assume values between 0 and 1, and higher values indicate higher probability that the observed scores and the expected scores represent the same population. Note that this study tests whether two samples are the same. Therefore, a higher value of p is preferred since p indicates how likely it is that the two sets of scores are equivalent. Hence, this approach differs from the one used

in many studies that test whether two samples are different from each other in order to determine, for instance, whether two different teaching methods or medications yield different results. Such studies, therefore, seek a low value of *p*, typically one below 0.05 (Hartshorn 2015).

- $\bar{x}_D$ = mean of the differences between the observed and calculated scores. For each student, the difference between their observed exam score and predicted exam score is calculated, and $\bar{x}_D$ is the average of these differences. Consequently, it is also the difference between the average of the observed exam scores and the average of the expected exam scores. The closer it is to zero, the better. Positive values indicate that the equation generally overestimates the expected scores, and negative values indicate underestimation. Negative values are preferred to positive ones because it is probably better to urge a student who is not in danger of failing the exam to do even better than to miss a student who is in danger of failing.
- s_D = standard deviation of the differences between the observed and expected scores. The lower it is, the better it is because that indicates lower variability in the predictions.
- min_D = the lowest value of the differences between the observed and expected scores. It is the longest distance between an underestimated score and the actual score. The closer it is to zero, the better.
- max_D = the highest value of the differences between the observed and expected scores. It is the longest distance between an overestimated score and the actual score. The closer it is to zero, the better.
- mad_D = the mean absolute deviation of the differences between the observed and expected scores. It is measured from $\bar{x}_D$. The lower it is, the better it is for the same reason as with the standard deviation.
- $<s_D$ = the number of expected scores that are within a distance of one standard deviation from $\bar{x}_D$. If the distribution of the difference is normal, it should amount to about 68% of the number of informants. In any case, the higher it is, the better it is because it indicates that more predictions are closer to their target.
- $>2s_D$ = the number of expected scores that are further from $\bar{x}_D$ than two standard deviations. If the distribution of the difference is normal, it should amount to about 5% of the number of informants. In any case, the lower it is, the better it is because missing the target by more than two standard deviations indicates an all too low predictive power of the equations.
- $<\pm 3$ = the number of expected scores that are within 3 percent points of their targets. 3 percent points has been chosen as the limit because the average interval of the grades is 6 percent points. Hence, 3 percent points is the largest deviation which, at the same time, is the least likely one to cause a target miss which amounts to a change in grade. The higher, the better.

For the datasets, the following abbreviations are used.

- *Aggr*. is the dataset consisting of the aggregate scores of the course-initial test and the home assignments.
- *Plain* is the dataset that consists of the plain scores of the individual grammar topics.
- *Ord*. is the dataset that consists of the scores of the individual grammar topics which are differentiated on an ordinal scale.

Prop. is the dataset that consists of the scores of the individual grammar topics which are differentiated on a proportional or interval scale, cf.

- Table 3.

5.1 Multiple linear regression

The tables below summarize the results of the multiple linear regression analysis. First, the equations were checked against their own samples in order to see how they perform under the most optimal circumstances (Table 4 and Table 5). Then, their performance was checked against the other sample, which was the very purpose of this study (Table 6 and Table 7). The most favorable value of a given metric is highlighted in each column. The favorable values do not necessarily coincide. Hence, selecting the best performing equation is not a trivial matter. The *<±3* values are of course weighted high because they express how many predictions come within the preferred range around the target. However, the spread of the predictions is also very important since it is desirable that the predictions do not miss their target by too large a margin.

Table 4: Equations for Year 2014 checked against Year 2014

mean of observed scores = 70.39, number of informants = 60												
dataset	initial test	$\bar{x}_E$	r_{EO}	*p*	$\bar{x}_D$	s_D	min_D	max_D	mad_D	$<s_D$	$>2s_D$	<±3
aggr.	incl.	70.47	0.76	0.94	0.08	8.58	-16.56	23.16	6.66	73%	5%	33%
	excl.	70.52	0.75	0.91	0.13	8.75	-16.08	24.20	6.72	72%	7%	30%
plain	incl.	70.44	**0.89**	0.95	0.05	**6.02**	-11.81	18.48	**4.81**	68%	**3%**	40%
	excl.	70.43	**0.89**	0.95	0.05	6.07	**-11.16**	18.77	4.83	70%	**3%**	38%
ord.	incl.	70.40	0.88	**0.99**	0.01	6.32	-13.25	17.25	4.91	72%	7%	37%
	excl.	70.40	0.87	0.98	0.02	6.54	-12.38	18.32	5.02	**75%**	5%	37%
prop.	incl.	70.36	0.88	0.98	-0.03	6.38	-12.92	**16.91**	4.87	65%	7%	42%
	excl.	**70.38**	0.87	**0.99**	**-0.01**	6.60	-12.76	17.81	4.94	72%	5%	**43%**

Table 5: Equations for Year 2015 checked against Year 2015

mean of observed scores = 64.50, number of informants = 79												
dataset	initial test	$\bar{x}_E$	r_{EO}	*p*	$\bar{x}_D$	s_D	min_D	max_D	mad_D	$<s_D$	$>2s_D$	<±3
aggr.	incl.	64.59	0.85	0.93	0.08	8.77	-23.25	23.06	6.55	73%	8%	**34%**
	excl.	64.55	0.81	**0.96**	0.05	9.72	-28.11	22.85	7.71	**78%**	9%	24%
plain	incl.	**64.45**	**0.92**	0.95	**-0.05**	**6.61**	**-15.60**	15.23	**5.33**	66%	5%	32%
	excl.	64.45	0.91	0.95	-0.06	7.08	-16.97	**13.87**	5.83	66%	**3%**	27%
ord.	incl.	64.37	0.91	0.86	-0.14	6.91	-17.59	17.87	5.49	71%	4%	**34%**
	excl.	64.36	0.90	0.86	-0.15	7.34	-21.97	17.33	5.75	68%	4%	33%
prop.	incl.	64.30	0.90	0.80	-0.21	7.28	-16.99	16.24	5.80	72%	8%	33%
	excl.	64.27	0.89	0.79	-0.23	7.76	-20.96	16.55	6.08	67%	5%	33%

Looking at the results of the "self-test" of the equations, it is evident that the equations gained from the regressions based on the individual grammar topics are superior to the ones which are gained from

the regression based on the aggregate scores. It is also clear that the equations that also take the course-initial test into account tend to perform better than the corresponding ones without the course-initial test. On the other hand, it does not seem to pay off to invest the extra computational effort into the differentiation of the questions within the grammar topics, although it is not detrimental either.

The distribution of the differences is close to normal distribution. The $<s_D$ values are around 68%, and the $>2s_D$ values do not typically exceed 5%. The $<\pm 3$ values are around 33%, which is somewhat mediocre.

Table 6: Equations for Year 2015 checked against Year 2014

mean of observed scores = 70.39, number of informants = 60												
dataset	initial test	$\bar{x}_E$	r_{EO}	p	$\bar{x}_D$	s_D	min_D	max_D	mad_D	$<s_D$	$>2s_D$	$<\pm 3$
aggr.	incl.	67.67	**0.72**	0.03	-2.72	**9.36**	**-24.74**	**21.79**	**7.03**	72%	7%	**28%**
	excl.	67.45	0.70	0.02	-2.94	9.48	-24.93	27.22	7.06	67%	8%	23%
plain	incl.	67.81	0.61	0.10	-2.57	11.67	-33.54	28.24	9.29	67%	5%	20%
	excl.	67.85	0.57	0.12	-2.54	12.24	-35.13	28.19	9.45	**73%**	7%	17%
ord.	incl.	69.27	0.57	0.51	-1.11	12.88	-35.97	29.98	10.08	67%	5%	22%
	excl.	69.48	0.51	0.62	-0.90	14.03	-39.78	36.23	10.91	68%	5%	18%
prop.	incl.	69.57	0.59	0.63	-0.82	12.95	-36.06	32.90	10.44	67%	**3%**	12%
	excl.	69.81	0.52	**0.76**	**-0.57**	14.28	-40.03	41.50	11.18	72%	**3%**	15%

Table 7: Equations for Year 2014 checked against Year 2015

mean of observed scores = 64.50, number of informants = 79												
dataset	initial test	$\bar{x}_E$	r_{EO}	p	$\bar{x}_D$	s_D	min_D	max_D	mad_D	$<s_D$	$>2s_D$	$<\pm 3$
aggr.	incl.	67.71	**0.82**	0.01	3.21	**9.62**	**-18.55**	31.74	**7.45**	**76%**	8%	**32%**
	excl.	67.79	0.79	0.01	3.28	10.30	-20.20	31.46	8.11	71%	6%	29%
plain	incl.	65.05	0.70	**0.69**	**0.54**	12.12	-43.27	**30.27**	9.01	75%	5%	24%
	excl.	65.27	0.72	0.57	0.77	11.77	-41.52	30.35	8.77	72%	5%	**32%**
ord.	incl.	65.34	0.65	0.58	0.84	13.21	-37.35	36.93	9.93	68%	5%	24%
	excl.	65.96	0.72	0.28	1.45	11.87	-35.75	33.93	8.86	68%	**4%**	29%
prop.	incl.	67.31	0.66	0.07	2.81	13.36	-36.52	38.82	9.97	71%	6%	30%
	excl.	67.92	0.72	0.01	3.42	12.11	-34.96	35.69	9.04	70%	5%	23%

Several differences emerge from Table 6 and Table 7. First, the correlations in the case of controlling the equations against the group of informants from the other year are much lower than the ones gained when checking the equations against the year for which they were developed. This is also reflected in the much lower level of statistical significance and the higher values of the difference between the

expected and observed scores. Nevertheless, in most cases the correlations still reach an acceptable level of around 0.7.

Second, excluding the course-initial test does not make so large a difference between the members of the equation pairs as when evaluating the equations against their own year. In several cases, the equations without the course-initial test are in fact more precise than the ones incorporating it.

Last, the equations based on the individual grammar topics are not in any obvious way superior to the equations based on the aggregate scores as they are when being checked against their own years. Also, the patterns of the equations from the two years are quite different from each other. The equations from 2014 seem to agree better with the informants from 2015 than vice versa. However, the distribution of the differences is still close to normal.

As it turns out, the two samples are indeed different from each other. The *p* value from a *t*-test comparing the exam scores from the two years equals 0.0228, which suggests that the two samples are statistically significantly different from each other. It explains why the results in Table 6 and Table 7 are so different. Since the 2014 equations do better on their own sample than the 2015 equation on theirs, it is probably not surprising that this difference also emerges in a cross-sample comparison. It has likely to do with the fact the sample from 2014 is more homogenous than the sample from 2015 as their standard deviations are consistently lower (Table 8).

Table 8: Standard deviations of the samples

	course-initial test	home assignment 1	home assignment 2	home assignment 3	exam
2014	18.10	9.06	12.96	12.37	13.35
2015	18.82	11.24	14.04	15.16	16.76

It is, however, not a disadvantage that the two samples have proved to be so different because it hints at what can be expected when an equation based on one sample is used on another sample that is rather different from the former sample. If the results in Table 6 and Table 7 can be deemed satisfactory enough, then there is a good chance that the equations developed in this project can be used on further samples with some confidence.

As for the satisfactoriness of the predictions, it must be noted that even though $\bar{x}_D$ is a magnitude larger in the cross-checking than when the equations are compared against their own sample, it can still be as low as 0.54. Even a value of -2.57 (Table 6) is an acceptable prediction. Despite the significant increase in $\bar{x}_D$, the number of predictions within the ±3-percent-point margin is still in the range of 20 to 30%. Thus, the equations seem to be reasonable tools to detect students that may be lagging behind.

5.2 Polynomial regression

Table 9 shows the results of the polynomial regression analysis.

Table 9: Results of the polynomial regressions

mean of observed scores = 67.05, number of informants = 139													
polyn-omial		forced through origin	$\bar{x}_E$	r_{EO}	*P*	$\bar{x}_D$	s_D	min_D	max_D	mad_D	$<s_D$	$>2s_D$	$<\pm3$
3rd degree		yes	91.34	0.51	0.00	24.29	45.67	-106.64	154.46	36.38	62%	**7%**	6%
		no	149.22	0.01	0.00	82.17	58.38	-75.20	254.70	44.80	35%	24%	0%
1st degree	with test	yes	96.38	**0.75**	0.00	29.33	**11.56**	**-7.14**	68.74	**8.87**	6%	71%	1%
		no	76.58	0.57	0.00	9.54	15.06	-29.50	61.60	11.98	62%	9%	15%
	w/o test	no	**63.75**	0.67	0.00	**-3.30**	14.22	-42.07	**45.93**	10.40	**77%**	**6%**	**20%**

Somewhat disappointingly, though not entirely unexpectedly, the 3rd degree polynomials did not fulfill the expectations, performing rather poorly. Many predictions missed the target by a very large margin even exceeding the absolute limits of 0 and 100. The polynomials could possibly perform better if there were more data points, i.e. the scores from more exercises during the course, to build on; however, that is not possible for the time being.

On second thought, it is not so surprising that the first degree polynomials did better than the 3rd degree ones despite the agreeable curve in Figure 1. When having so few data points, it is likely that a linear approximation does a better job than a higher order one. This might, however, change if there were more data points, as alluded to above. The 1st degree polynomial (last row in Table 9) which was not pegged to the origin and which did not take the course-initial test into account did in fact quite well, almost on par with the equations from the multiple linear regression analyses. When considering that the 1st degree polynomial has the advantage that it does not need a sample base and has minimal computational requirements, it is not an unattractive alternative to multiple linear regression despite its somewhat lower precision.

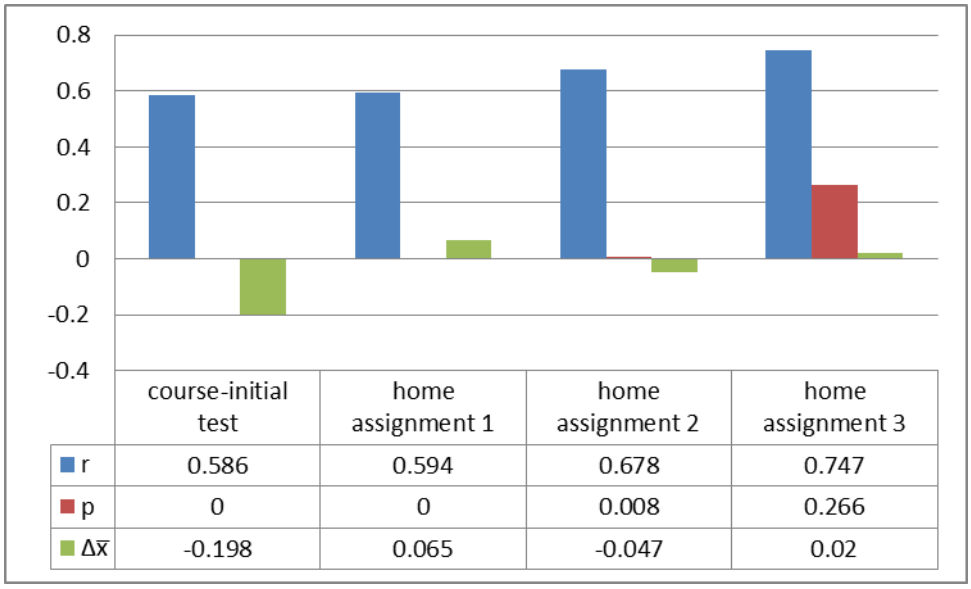
5.3 Correlation analysis

Figure 2 shows the results of the correlation analysis.

- *r* stands for the correlation coefficient between the exam scores and the respective exercise.
- *p* is the statistical significance of the correlations.
- $\Delta\bar{x}$ is the difference between the average of the exam scores and the average of the scores of the respective exercise. The former is subtracted from the latter. It has been scaled down by 100 so that it can be shown on the same scale as *r* and *p*.

The calculations are based on the combined sample of 2014 and 2015, which comprises 139 informants. The average exam score of the combined sample is 67.05.

Figure 2: Results of the correlation analysis



	course-initial test	home assignment 1	home assignment 2	home assignment 3
r	0.586	0.594	0.678	0.747
p	0	0	0.008	0.266
$\Delta\bar{x}$	-0.198	0.065	-0.047	0.02

As can be seen, the correlation with the exam increases steadily during the course. So does the probability of equality (*p*) between the exam scores and the scores of the respective exercises although it does not reach high. On the other hand, the gap between the exam scores and the scores of the exercises shows a steady decrease, though with fluctuation in its sign.

All these metrics suggest that the students are being prepared for the exam during the course, at least partly by the exercises they do. The calculations also suggest that the initial level of knowledge is not decisive for the exam score, and students with a low initial level can improve their knowledge sufficiently. This observation is also corroborated by the results of regression analyses. When the equations from the multiple linear regression analysis are cross-checked against the other sample, the equations incorporating the course-initial test do not perform noticeably better. It suggests that the results of the course-initial test are not good predictors of the exam scores. This suggestion is further strengthened by the 1st degree polynomials performing best when the course-initial test is not taken into account.

However, it does not mean that the course-initial test was taken in vain. It may have given a hint to the students about their standing and may have given them some motivation to make more effort, especially those who did not do well at the test.

6. Conclusion

The paper sought to answer two research questions. As for the question how well the home assignments prepare the students for the course-final exam, the study clearly shows that there is an increasing convergence between the students' exercise results and their exam results. It suggests that the exercises during the course have a beneficial effect for preparing the students for the exam. Of course, it does not guarantee that the existing exercises are the best possible ones; however, they certainly serve their purpose.

For the question how predictable the outcome of the grammar exam is, the results are promising. In 20 to 30% of the cases, the predictions are spot on, and within ±2 grades in 60 to 70% of the cases. It means that the teacher can confidently identify and – most importantly – notify those students who are in danger of failing the exam in advance. Even though this identification is possible only after the final home assignment has been evaluated, there is still some time until the exam so that the students notified can make extra effort to prepare for the exam.

The comparison of the prediction methods indicates that it is not necessary to differentiate between the questions of the individual grammar topics for the sake of the prediction, nor does it pay

off to increase the granularity of the scores. The basic aggregate scores are sufficient. There is furthermore close competition between the predictive equations developed through multiple linear regression analysis and the 1st degree polynomial regression, suggesting that the simplest prediction methods work just fine. All this means that putting the findings of this paper to practical use is no more complicated than setting up a simple Excel spreadsheet in which the teacher inserts a student's home-assignment scores and obtains the student's expected exam score immediately. The teacher and the student can then take action accordingly.

Finally, the study has also given an answer to an unasked, yet important question. Both the correlation and regression analyses suggest that the students' initial level of knowledge is not decisive for their exam score. In other words, also students with a limited understanding of grammar at the beginning of the course can learn it during the course; good exam grades are not reserved for the few who enter the university with substantial preunderstanding of grammar. This corresponds well with the conclusion that the home assignments support the students' learning adequately.

References

Boundless (2014). *Statistics*. Boundless. Kindle Edition.

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.

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

Elbro, Carsten & Hollis Scarborough (2003). "Early identification". In Terezinha Nunes & Peter Bryans (eds.), *Handbook of Children's Literacy*. Dordrecht: Kluwer Academic Publishers. 339-359.

Falls, Scott (2011). *Excel Formulas Revealed - Master Statistical Formulas in Microsoft Exce*l. Firefalls Publishing. Kindle Edition.

Hair, Joseph F., William C. Black, Barry J. Babin & Rolph E. Anderson (2010). *Multivariate Data Analysis - A Global Perspective*. Upper Saddle River, NJ: Pearson Education.

Harmon, Mark (2014). *Practical and Clear Graduate Statistics in Excel - The Excel Statistical Master*. Kindle Edition.

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

Hartshorn, Scott (2015). *Hypothesis Testing: A Visual Introduction to Statistical Significance*. Amazon: Kindle Edition.

Hartshorn, Scott (2017). *Linear Regression and Correlation*. Amazon: Kindle Edition.

Hjulmand, Lise-Lotte & Helge Schwarz (2012). *A Concise Contrastive Grammar of English*. Frederiksberg: Samfundslitteratur.

Madsen, Richard (2014). "Correlation between theoretical knowledge of grammar and performance in the production of written texts". In Lotte Dam & Rita Cancino (eds.), *Multidisciplinary Perspectives on Language Competence*. Aalborg: Aalborg University Press. 23-60.

Madsen, Richard (2015). "A statistical model of learning descriptive grammar". In Anna Bondaruk, Anna Bloch-Rozmej, Wojciech Malec, Ewelina Mokrosz & 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 (2017). *What is wrong with Grammar? Danish university students' difficulties with the acquisition of written English and theoretical grammar*. Aalborg: Aalborg University Press.

Madsen, Richard (2019). "Adaptive grading systems, or pros and cons of different ways of grading grammar exams" *Globe: A Journal of Language, Culture and Communication,* 9: 133-154.

Study Board of Language and International Business Communication, Aalborg University (2017). *Study regulations*. https://www.fak.hum.aau.dk/digitalAssets/415/415610_ba-siv17.pdf. Retrieved February 20, 2020.

Urdan, Timothy C. (2012). *Statistics in Plain English*. Routledge: Kindle Edition.