Eurasia Journal of Mathematics, Science & Technology Education, 2016, **12**(4), 975-989

A Quantitative Analysis of Uncertainty in the Grading of Written Exams in Mathematics and Physics

Hugo Lewi Hammer & Laurence Habib
Oslo and Akershus University College of Applied Sciences, NORWAY

•*Received 1 September 2015•Revised 18 November 2015 •Accepted 20 November 2015*

The most common way to grade students in courses at university and university college level is to use final written exams. The aim of final exams is generally to provide a reliable and a valid measurement of the extent to which a student has achieved the learning outcomes for the course. A source of uncertainty in grading students based on an exam is that such exams only consist of a limited number of exercises. We investigate the extent of this uncertainty by means of a statistical analysis of the results of 23 different examinations taken by 2788 students. The amount of uncertainty is substantial and typically ranges over three grades. Increasing the duration of the examination decreases the uncertainty, however.

Keywords: examination duration, grading, quantitative research, uncertainty, written exam

INTRODUCTION

Background to the study

Quality in higher education has been the focus of much policy development worldwide (Blanco-Ramírez & Berger, 2014), often in response to external incentives or to comply with norms that are considered legitimate (Vukasovic, 2013). This situation has resulted in increased focus on devising and implementing quality assurance systems in institutions of higher education (Westerheijden, Stensaker, Rosa, & Corbett, 2014). The notion of academic quality is multifaceted and has been described as "an inherently vague concept" (Wittek & Kvernbekk, 2011). It may also be noted that, when measuring the quality of education, the emphasis has shifted from educational inputs (what the teacher conveys and how) to learning outputs (what the students have achieved in terms of learning outcomes), as reported in Hughes (2013). In that respect, it is natural to consider the quality of assessment methods as an inherent part of the quality of an educational program (Boyas, Bryan, & Lee, 2012).

Examination results have been referred to as a form of "currency" (William, 1996; Simpson & Baird, 2013) that is dependent on trust in order to retain its status

Correspondence: Hugo Lewi Hammer,
Department of Computer Science, Oslo and Akershus University College of Applied Sciences, PO box 4 St. Olavs plass NO-0130 Oslo, Norway.
E-mail: hugo.hammer@hioa.no
doi: 10.12973/eurasia.2016.1240a

ISSN: 1305-8223

and value. In tertiary education, the primary purpose of final examination grades is to communicate a student's achievement to future employers and to other institutions to which the student might apply for further degrees. It is generally accepted that grades can be of critical importance to a student's future educational path and career, and that it is therefore crucial to ensure that they accurately reflect the student's proficiency level.

Issues relating to the reliability and validity of assessment have been the focus of attention in the literature on assessment. Reliability has been defined as "the repeatability of an assessment and its results" (Irwin & Hepplestone, 2012, 777). An assessment form can be said to be reliable if it is not affected by factors that lay outside the student's control, such as the background or the views of the examiner (Harlen, 2005). Changes in difficulty levels from one year to another, or from one semester to another, may also endanger the reliability of an assessment form (DeVellis, 2012).

Validity refers to the extent to which an assessment form "measures what it is designed to measure" (Russell, Elton, Swinglehurst, & Greenhalgh, 2006, 466). In that respect, an assessment form is only valid if it allows the students to demonstrate effectively whether and to what degree they have achieved the learning goals that were set for the course. A valid assessment is therefore one that prevents students from over-performing or under-performing compared with their actual level of mastery of the curriculum. Altogether, validity necessitates reliability, but reliability is not in itself sufficient to ensure validity.

Institutions of higher education throughout the world are currently under pressure to increase productivity due to budget cuts (Agasisti & Bonomi, 2014) and growing student numbers (Allais, 2014). In an educational climate where cost-efficiency is emphasized, institutions may feel pressure to reduce the duration of examinations in order to reduce the costs associated with remunerating invigilators, renting examination rooms, and compensating faculty members, teaching assistants or external examiners for marking the examination papers. In addition, institutions may face examination timetabling problems due to the limited number of weeks that can be allocated to examinations and because of growing student numbers (as suggested in, e.g., Mumford (2010) and Abdul-Rahman, Burke, Bargiela, McCollum, & Özcan (2014)).

Compared with other types of summative assessment, such as oral examinations, closed-book written examinations at the end of the module, semester, or academic year are relatively inexpensive. Computer-based approaches (as described in, e.g., Delen (2015) and Kuo, Daud, & Yang (2015)) are another inexpensive approach, although they are much less used than written exams. Written examinations are regarded as particularly suitable for testing students' learning outcomes in mathematics and other science subjects (Davis, Harrison, Palipana, & Ward, 2005). Relatively little attention has been devoted, however, to the reliability and validity of

State of the literature

- There is a long tradition of using statistical approaches to analyze the reliability and validity of written exams, see, e.g., Lord (1952, 1953), Lord and Novick (1968).
- More recent advances: Item Response models (Hambleton, Swaminathan, & Rogers, 1991; Lord & Novick, 1968; Lord, 1980) and the Generalized Partial Credit model (Muraki & Bock, 2002; Muraki, 1997)
- References focusing on the number of exercises that should be included in a test or exam to alleviate the challenges of uncertainty in grading: Bird and Yucel (2013) and Burton (2006).

Contribution of this paper to the literature

- We have not found any paper analyzing real exam correction data to measure uncertainty in the grading of written exams in mathematics and physics.
- The most likely reason is that none of the traditional models, such as the Item Response model or the Generalized Partial Credit model, are applicable to such data. Instead, we construct a suitable model based on the less common Beta Regression framework.
- We demonstrate that exam correction data are a very valuable source of information for measuring uncertainty in the grading of written exams in mathematics and physics. Our results show that the uncertainty is substantial and typically ranges over three grades.

written examinations. Interestingly, the literature on assessment seems to be more concerned with the validity of other assessment forms, such as practical examinations (Vu et al., 2006), modified essay questions (Palmer, Duggan, Devitt, & Russell, 2010), or portfolio assessment (Admiraal, Hoeksma, van de Kamp, & van Duin, 2011).

There are a few notable exceptions, however, for example the works of Bird and Yucel (2013) and Burton (2006). The latter suggests that the optimal length of an academic test consisting of short-answer and multiple-choice questions with dichotomous scoring (either 0 or 1) might be around 300 questions or test items. This is in order to allow for different levels of difficulty in the questions, unevenness of knowledge among the students taking the test, and the possibility that some questions may be badly phrased, while other questions may be so similar to the textbook material that students can answer them correctly more from memory than by reasoning. He also points out that testing more than two separable facts in one dichotomously scored test item provides an additional level of uncertainty and recommends avoiding the use of such "double questions" (p. 576).

There is a long tradition of using statistical approaches to analyze the reliability and validity of written exams. Foundational works such as Lord (1952) and Lord (1953) have highlighted the need to differentiate between ability scores, which are test-independent, and observed scores and true scores, which are test-dependent. Other works, such as Lord and Novick (1968), have described classical test theory as relying on the assumption that test scores are the result of a combination of true scores and measurement error.

The study described in this article aims to provide insights into the reliability and validity of written exams in mathematics and physics. To that end, we present an extensive analysis of uncertainty in the grading of written exams. Such exams typically consist of 10 to 20 exercises from different parts of the curriculum, and the reliability is affected, among other things, by the number of exercises included. We analyze the reliability of such exams using a quantitative approach based on an extensive dataset consisting of the marking of 34 800 examination answers from 2788 students based on exams from two universities and one university college in Norway. We analyze the data using a Generalized Linear model (Dobson & Barnett, 2008). Generalized Linear models have been applied extensively in educational measurement or educational assessment through models such as Item Response models (Hambleton, Swaminathan, & Rogers, 1991; Lord & Novick, 1968; Lord, 1980) and the Generalized Partial Credit model (Muraki & Bock, 2002; Muraki, 1997). It is worth noting, however, that all these models assume that the test scores are discrete (e.g., right/wrong). This suggests that traditional assessment models cannot be used to shed light on data material where the scores are continuous, as is the case in our study. The analysis in this article is thus based on a less common Generalized Linear model called Beta Regression.

To the best of our knowledge, there is little published research that takes a quantitative approach to analyzing the reliability and validity of written exams. We assume that the reason for this is that it is not possible to analyze continuous data using the traditional statistical assessment models described above. Our decision to use beta regression may therefore represent a significant contribution to the field of assessment, since it provides new insights into the reliability and validity of written exams.

Examples

In order to ensure both the reliability and validity of exams in mathematics or physics, such exams must include a sufficient amount of exercises to test the students' actual level of mastery. The following example could be used to illustrate this claim. Let us consider two students with very different levels of mastery of the course curriculum, which consists of 10 main parts. If student A masters only one of the ten parts and student B masters nine of the ten, an examination with only one exercise aimed at testing just one part of the curriculum might give a totally erroneous picture of the students' actual level of mastery. If the exercise happens to be on the one part of the curriculum that student A masters, he or she will get a good grade, which does not reflect his or her actual level of mastery of the curriculum. Conversely, if the exercise happens to be on the one part of the curriculum that student B does not master, he or she will be awarded a poor grade that does not reflect his or her level of mastery either. In order to reduce the random effect of luck (or lack thereof) on examination scores and thereby increase their validity and reliability, it is necessary to ensure that each examination consists of a sufficient amount of exercises.

A second example may further illustrate the problem. We assume that an exam consists of an equal amount of very difficult, difficult, easy, and very easy exercises from different parts of the curriculum. For an average student, we assume that the probabilities of the student managing exercises on the different levels of difficulty are 0.2, 0.4, 0.6, and 0.8, respectively. We assume that the time allotted per exercise is 15 minutes, which is typical for traditional exams in mathematics and physics. The exam score will be the mean of the scores for each exercise. Figure 1 shows the probability of different exam scores for the student in this simple binomial model. The exam scores are rescaled to a 0 to 100 scale. We see that, for a one-hour exam (four exercises), there is a probability of approximately 4% that the student will get

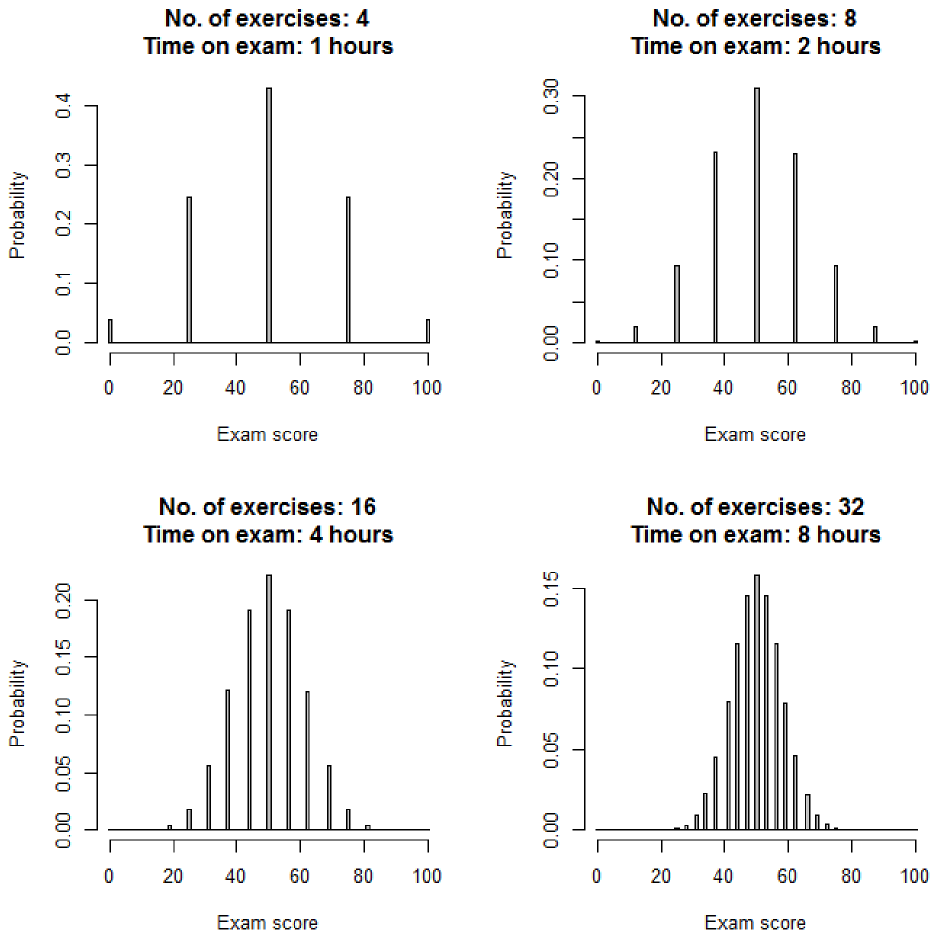

Figure 1. Probability of different exam scores in a binomial exam model

all the answers wrong and an equal probability that the student will get all the answers correct, which means that the reliability of such a short exam is very poor. In the case of four-hour and eight-hour exams, the possible exam results are spread over several grades as well, which means quite poor reliability. In the rest of this paper, a similar analysis will be performed where the uncertainty (lack of reliability) is estimated based on the data from the marking of the 34 800 examination answers.

MATERIAL AND METHODS

Exam correction data

Our analysis was based on the marking of 23 exams for introductory courses in mathematics, statistics, and physics from the University of Oslo, the Norwegian University of Science and Technology, and Oslo and Akershus University College of Applied Sciences. The material consists of marks awarded to 2 788 different students for 301 different exercises. The marking of each exercise for all the students in all the exams will be used in the analysis, ending up with a total of 34 800 observations. For each exam, the marks (scores) for the exercise answers were normalized to the [0, 1] interval, where completely wrong and completely correct answers were awarded zero and one point, respectively.

The characteristics of the dataset were as follows. Each exercise in the data material required the student to perform some kind of calculations, i.e., no multiple choice exercises where the student could guess the answer. All exams were traditional written exams using pen and paper.

The duration of the exams varied between three and five hours. The time allotted to solving each exercise varied between the different exams, ranging from 12 minutes to 18 minutes. For exercises where the students were given a long time per exercise, the exercises typically consisted of many subtasks or longer computations.

Methodological issues

The markings (scores) of exercises from earlier written exams are an extremely useful source of information for measuring the reliability and validity of written exams, as will be seen in the results section. Quite surprisingly, we have not found any research papers that take advantage of this valuable source of information. The most likely explanation is that the data are available in a format that does not easily lend itself to analysis. In this article, we show that Beta Regression, a type of Generalized Linear model (Dobson & Barnett, 2008) is a suitable choice. A motivation for and detailed description of the statistical model is provided below.

Statistical model

In this section, we describe a statistical model that quantifies the amount of uncertainty in the grading of written exams in mathematics and physics. The statistical model has much in common with Item Response models and Generalized Partial Credit Interval, but our model differs from these models in that the responses (test scores) are not discrete, but continuous on a limited interval. Naturally, if the uncertainty in grading is high, the reliability and validity of the exam will be low. Let M quantify the level of mastery for a student taking an exam. A student with a high level of proficiency in the subject will have a large value of M, while a student with a low level of proficiency in the subject will have a small value. Further, let D quantify the level of difficulty of an arbitrary exercise in an exam. An easy exercise will have a low value of D, while a difficult exercise will have a large value of D. Let $S_1, S_2, \dots, S_n$ denote the scores for a student for the different exercises in an exam. We assume that each score is given on the [0, 1] interval, where a completely wrong answer results in the score zero, a completely correct answer in the score one and a partly correct answer somewhere in between. Let S_E denote the resulting exam score (grade) for this student based on the exercise scores $S_1, S_2, \dots, S_n$. The most common way to compute the exam score, S_E, is to take the average of each exercise score S_i and multiply by 100

$$S_E = \frac{100}{n} \sum_{i=0}^{n} S_i \tag{1}$$

For an exam to have high reliability and validity, the uncertainty of the exam score, S_E, must be low. The main source of uncertainty in the exam score is that, if a student is given many exercises of the same level of difficulty, D, the student will by chance alone get some exercises correct, some wrong, and some partly correct. If the student is lucky, an exam will consist of many exercises that the student, by chance, is able to solve. If the student is unlucky, the exam will consist of many exercises that the student, by chance, is not able to solve. Recall the two examples at the end of the introduction. The best way to reduce this source of randomness in exam score, S_E, is to include many exercises that are well-suited to testing different parts of the curriculum. Since the exam score is typically the average of the exam scores (Equation (1)), by the law of large numbers, the exam score will approach the student's true level of mastery, m, when the number of exercises increases.

As described above, a student who is given many exercises of the same level of difficulty, D, will simply by chance get some exercises correct, some wrong, and some partly correct. Let $p(s; m, d)$ denote a probability distribution that summarizes this property. More specifically, $p(s; m, d)$ is the probability distribution of exercise scores, S, a student with a level of mastery M will be awarded for an exercise of difficulty level D. If this probability distribution is narrow (small variance), the uncertainty in exercise scores $S_1, S_2, \dots, S_n$ is small and the resulting uncertainty in exam score, S_E, will be small (recall equation (1) and the law of large numbers). If the distribution $p(s; m, d)$ is wide, the uncertainty in exam score, S_E, will be large. We also expect that, if a student has a high level of mastery M or the exercise is easy (low value of D), the distribution will shift toward high values of exercise scores, and shift toward low values if the student has a low level of mastery or the exercise is difficult.

We estimate the probability distribution $p(s; m, d)$ using a regression model where the exercise score S is the dependent variable and M and D are the independent variables, modelled as random effects. The distributions for level of mastery (M) of each student, level of difficulty (D) of exercises, and the relation between M, D, and S are estimated using the marking (scores) of the 34 800 exercises in the data material. Traditional statistical assessment models assume that the response is binomial. Since the exercise scores $S_1, S_2, \dots, S_n$ represent continuous variables on the [0, 1] interval in our data, the binomial regression models cannot be used. We instead used the less common alternative of assuming that the exercise scores are outcomes from a beta distribution. The beta distribution is a highly flexible continuous distribution on the [0, 1] interval depending on the choices of the model parameters, as shown in Figure 2. It is thus an ideal distribution for modelling the exercise scores $S_1, S_2, \dots, S_n$. The exercise scores are typically distributed with "U"–shapes like the black curves in Figure 2, since it is most common to score answers to exercises as either completely wrong (zero points) or completely correct (one point) (see Figure 3). A more detailed description of the beta regression model is provided in the Appendix.

RESULTS

As described above, the data material consists of the marking (score) of each exercise for all the students in the 23 exams, resulting in a total of 34 800 scores. Figure 3 shows a histogram of all these scores.

We see that the most common scores are, as expected, zero and one, but also the scores 1/6, 1/4, 1/3, 1/2, 2/3, 3/4, and 5/6 are used to some extent.

We now fit 23 beta regression models, one for each exam.

We start by showing results from one out of the 23 exams, as representative of the results of all the 23 exams. The estimated values for the parameters in the regression model are shown in the Appendix (Table 2).

Figure 4 shows the distribution of exercise scores, $p(s; m, d)$, for different levels of difficulty for a student who on average scores 50 out of 100 points in the exam

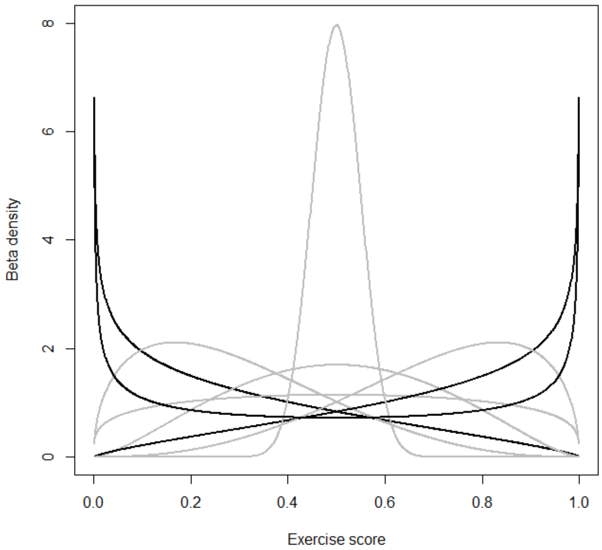

Figure 2. Beta distribution for a variety of values of the shape parameters (a, b)
Note: *The black curves show typical distributions of scores on exam exercises.*

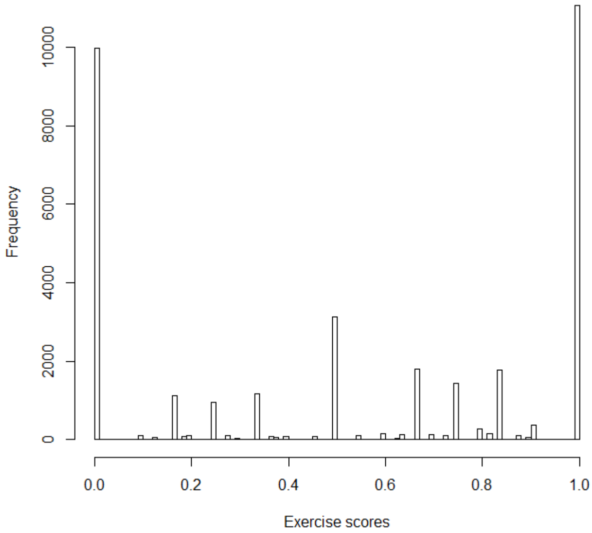

Figure 3. Distribution of scores for all the answers in the 23 exams

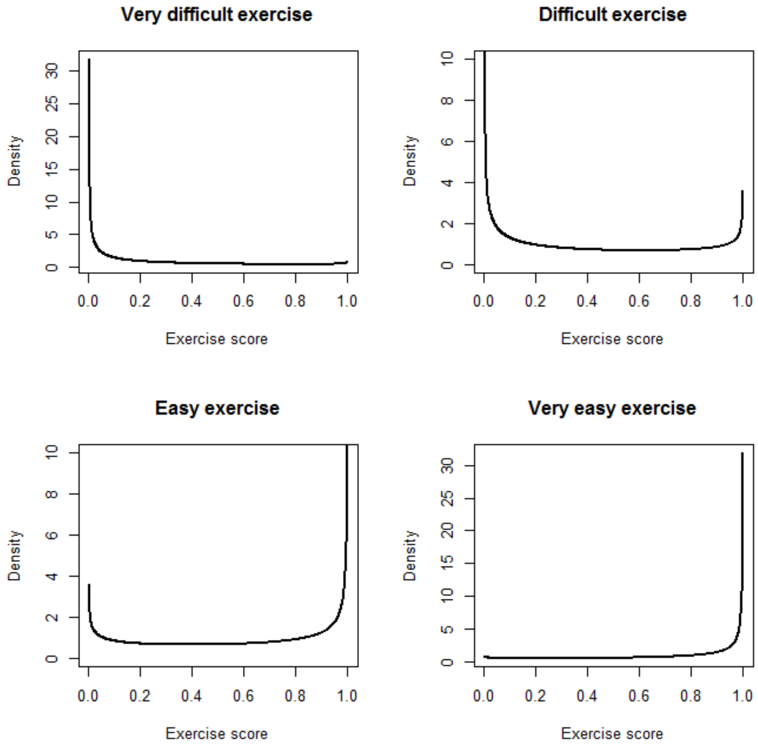

Figure 4. Distribution of exercise scores, p(s;m,d), for an average student for exercises of varying levels of difficulty

(average level of mastery M). Easy and very easy exercises are represented by exercises being one and two standard deviations easier than average exercises. Similarly, difficult and very difficult exercises are represented by exercises being one and two standard deviations more difficult than average exercises. We see that the distributions acquire the characteristic "U" shape, which is as expected since the most common exercise scores in the data material are zero and one (Figure 3).

Now, suppose that the average student faces an exam with an equal amount of very difficult, difficult, easy, and very easy exercises. For the exam that we will now study, the time per exercise was 15 minutes, which means that a four-hour exam consists of 16 exercises, four exercises on each level of difficulty. The exam score is computed from the exercise scores using equation (1). The distribution of possible exam scores for the student is shown in Figure 5. As expected, the uncertainty in the exam score is reduced as the number of exercises in the exam is increased (recall Equation (1) and the law of large numbers). Thus, the reliability and validity of the exam increases when the number of exercises increases. From Figure 5, we see that almost all possible exam scores that the average student can get for a four-hour exam fall between 30 and 70 points.

Figure 6 shows the same as Figure 5, but for a student who is two standard deviations stronger than an average student. By comparing Figures 5 and 6, we make the interesting observation that the uncertainty in the exam score is smaller for strong students than for average students.

We now present results for all the 23 exams. As a measure of uncertainty in exam scores, we use the difference between the 95% and 5% quantile in the distribution of exam scores. For example, for the four-hour exam in Figure 5, the 95% and 5% quantiles are 63.2 and 36.9, respectively, resulting in a difference of 26.3 points. As described in the methods section, the time allotted to solving an exercise varies between exams (12 to 18 minutes). A comparison based on the number of exercises is therefore not correct. An exam with a few exercises but with little variability in

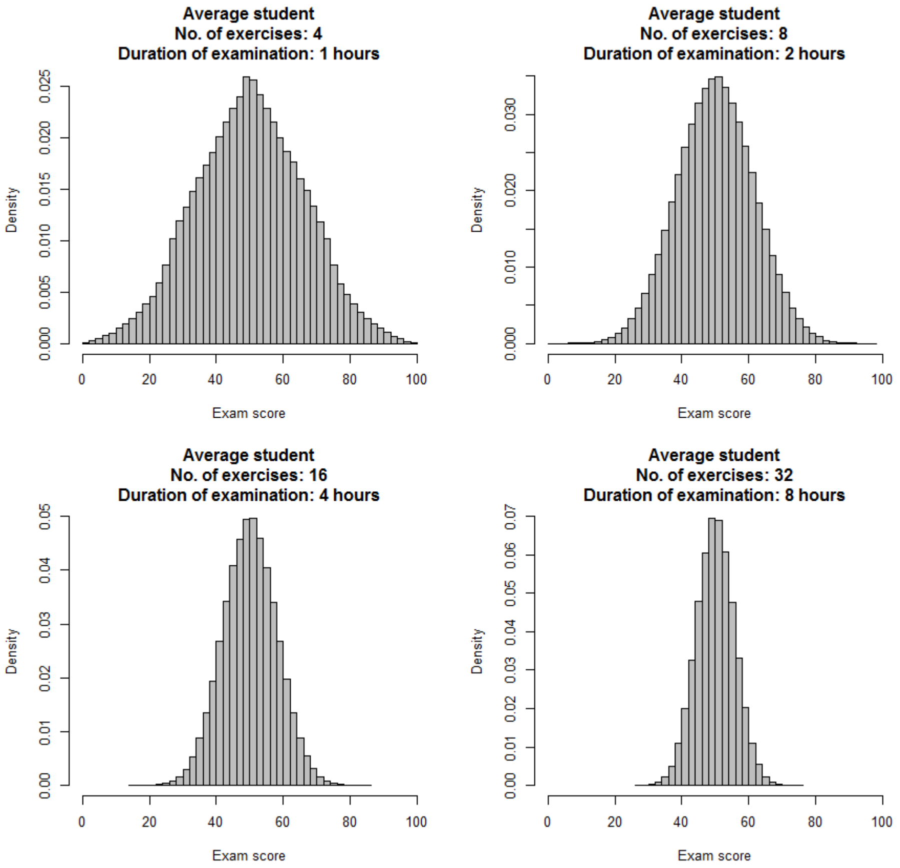

Figure 5. Distribution of exam scores for an average student

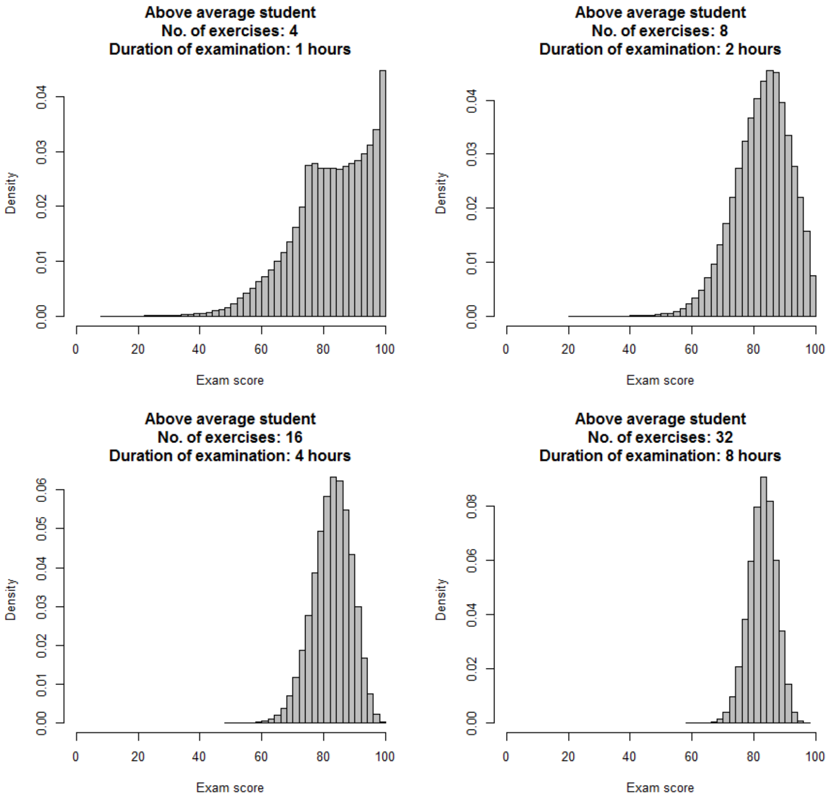

Figure 6. Distribution of exam scores for a strong student

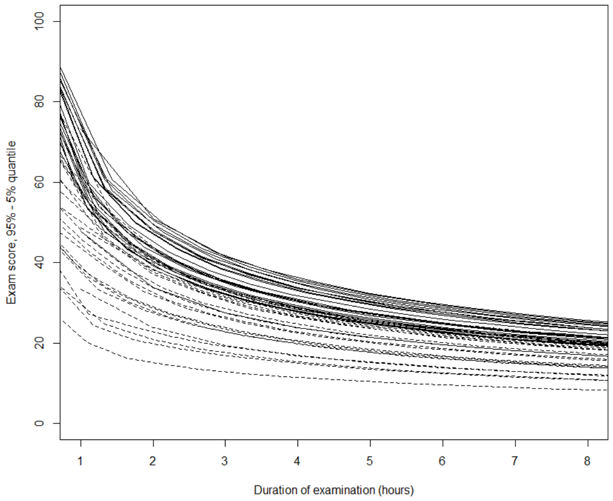

Figure 7. Uncertainty in exam scores as a function of duration of exam

the exercise scores can be better than an exam with many exercises and high variability. Since we know the time allotted per exercise for the different exams, we can re-compute from the number of exercises given to the duration of examination, and compare this to the uncertainty in exam scores. The results are presented in Figure 7 as described below.

Figure 7 shows the relationship between uncertainty in the exam score (the difference between the 95% and 5% quantile as described above) and the duration of examinations for all the 23 exams. We have two curves (dashed and solid) for each exam, representing cases with little and much uncertainty in the exam score. The main contribution to the varying uncertainty (difference between the 95% and 5% quantile) is the level of mastery of the students. There is less variability in exam scores for weak and strong students compared to average students (recall Figures 5 and 6). The uncertainty from the estimation of the true regression parameters is also included. As expected, in Figure 7, we see that the uncertainty is reduced when the duration of examinations increases (recall Equation (1) and the law of large numbers). For example, in a two-hour exam, the uncertainty for an average student potentially reaches above 50 points (out of 100), while in a four-hour exam, the uncertainty is rarely above 35 points and, for a six-hour exam, rarely above 25 points. For strong and weak students, the uncertainty is rarely above 40, 25, and 20 points, for two-hour, four-hour, and six-hour exams, respectively. We see some differences between the 23 exams, but overall the different exams have more or less the same amount of uncertainty in exam scores.

DISCUSSION AND CONCLUSION

The analysis shows that there is substantial uncertainty in grading written exams due to the limited duration of examinations. This means that the reliability and validity of written exams in mathematics and physics are critically low. By increasing the duration of examinations, the uncertainty will decrease and the reliability and validity improve. From Figure 7, however, we see that the reduction in uncertainty is less when we go from a two-hour to a four-hour exam compared to going from a four-hour to a six-hour exam.

The conversion of an exam score on the interval [0, 100] to specific grades varies a lot around the world, but the ECTS system (A-F) with conversions as shown in Table 2 is very common (Radboud University Nijmegen, 2011). For all international grading systems, the interval for each grade is typically between 5 and 20 points wide (except for the interval for ‘fail’). This means that the uncertainties documented in Figure 7 span several grades. For example, for the grading systems in Table 1, for a four-hour exam, an average student can be awarded all grades between F and C, while a strong student can be awarded all grades between C and A, on a purely chance basis. This means that the reliability and validity of such written exams is low.

The analysis in this paper confirms that increasing the length of examinations has a significant effect on reducing the amount of uncertainty in marking. Such results suggest that institutions should strive to use as long a duration as practically possible for written exams. Fatigue as a result of the long duration of an examination may be an issue, but previous research on examinations in other subject areas documented that performance increased with examination length (Jensen, Berry, & Kummer, 2013; Ackerman & Kanfer, 2009).

It can be noted that the results from our research were obtained by studying examinations where the various exercises covered as much of the curriculum as possible (typically, each exercise would be used to test the student’s mastery of a different area of the curriculum). If, for any reason, an examination is designed in such a way that it only aims to test parts of the curriculum (for example, if it includes several exercises that are related to the same part of the curriculum, and no exercises that are related to other parts of the curriculum), then increasing the length of the examination might not result in a decrease in marking uncertainty. In such cases, a longer examination might neither increase its reliability nor its validity, as it would be based on a biased sample of curriculum parts.

Table 1: Typical conversions to ECTS grading system (A-F)

Grade	Score intervals 1	Score intervals 2
A	90 – 100	90 – 100
B	80 – 89	80 – 89
C	70 – 79	60 – 79
D	60 – 69	50 – 59
E	50 – 59	40 – 49
F	0 – 49	0 – 39

It can also be noted that increasing the length of an examination will not contribute to reducing marking uncertainty if the examination is not designed to test mastery levels in a time-efficient way. In other words, increasing the length of an examination solely by including lengthy and tedious calculations in the exercises will not increase its reliability or its validity. In order to reduce uncertainty, it is necessary to design examinations in such a way that the time that students spend on answering questions is used as effectively as possible. For example, when an examination question only aims to test the students’ mastery of recalling facts, a multiple-choice form may be a better alternative than a lengthy exercise.

It can be inferred from the data and from our analysis that there is generally a large degree of uncertainty associated with using a summative assessment of one subject as an indicator of a student’s level of mastery of the curriculum in that subject. It is therefore unlikely that one grade will provide an accurate picture of a student’s abilities. In order to reduce this uncertainty, it is necessary to have access to a larger number of examination grades. Typically, a student takes between 25 and 75 exams within the framework of an educational program, and, because of the law of large numbers, the average grade based on all the individual course grades will normally reflect the student’s ability with much less uncertainty than an individual grade.

Of course, there are other possible challenges to the validity and reliability of an average grade based on several exams, and they could be the subject of further research. Such challenges might include differences in strictness levels from one examiner to another, from one subject area to another, and from one institution to another. Another challenge may be that some examiners and institutions use norm-referenced grades (i.e., grades that to a greater extent reflect where the examination paper stands in comparison with the level of the other examination papers), rather than criterion-referenced grades (i.e., grades that reflect the intrinsic quality of the paper, independently of the rest of the group). Although this practice has been pinpointed as unethical (Sadler, 2009), it is commonly used in various educational settings, for example in order to prevent “grade inflation” (Cliffordson, 2008). It is therefore important that further research encompassing a broad variety of examinations and examination results ascertains the degree of integrity of the grading systems, i.e., the extent to which they are criterion-based rather than norm-based.

REFERENCES

Abdul-Rahman, Syariza, Edmund Burke, Andrzej Bargiela, Barry McCollum, and Ender Özcan. 2014. "A constructive approach to examination timetabling based on adaptive decomposition and ordering." *Annals of Operations Research* 218 (1):3-21. doi: 10.1007/s10479-011-0999-8.

Ackerman, Phillip L., and Ruth Kanfer. 2009. "Test Length and Cognitive Fatigue: An Empirical Examination of Effects on Performance and Test-Taker Reactions." *Journal of Experimental Psychology: Applied* 15 (2):163-181. doi: 10.1037/a0015719

Admiraal, Wilfried, Mark Hoeksma, Marie-Therese van de Kamp, and Gee van Duin. 2011. "Assessment of Teacher Competence Using Video Portfolios: Reliability, Construct Validity, and Consequential Validity." *Teaching and Teacher Education: An International Journal of Research and Studies* 27 (6):1019-1028. doi: 10.1016/j.tate.2011.04.002

Agasisti, Tommaso, and Francesca Bonomi. 2014. "Benchmarking universities' efficiency indicators in the presence of internal heterogeneity." *Studies in Higher Education* 39 (7):1237-1255. doi: 10.1080/03075079.2013.801423.

Allais, Stephanie. 2014. "A critical perspective on large class teaching: the political economy of massification and the sociology of knowledge." *Higher Education* 67(6):721-734. doi: 10.1007/s10734-013-9672-2.

Bird, Fiona L., and Robyn Yucel. 2013. "Improving marking reliability of scientific writing with the Developing Understanding of Assessment for Learning programme." *Assessment & Evaluation in Higher Education 38*(5):536-553. doi: 10.1080/02602938.2012.658155.

Blanco-Ramírez, Gerardo, and Joseph B. Berger. 2014. "Rankings, accreditation, and the international quest for qualityOrganizing an approach to value in higher education." *Quality Assurance in Education: An International Perspective* 22 (1):88-104. doi: 10.1108/QAE-07-2013-0031.

Boyas, Elise, Lois D. Bryan, and Tanya Lee. 2012. "Conditions affecting the usefulness of pre- and post-tests for assessment purposes." *Assessment & Evaluation in Higher Education* 37 (4):427-437. doi: 10.1080/02602938.2010.538665.

Burton, Richard F. 2006. "Sampling Knowledge and Understanding: How Long Should a Test Be?" *Assessment & Evaluation in Higher Education* 31 (5):569-582. doi: 10.1080/02602930600679589

Cliffordson, Christina. 2008. "Differential Prediction of Study Success across Academic Programs in the Swedish Context: The Validity of Grades and Tests as Selection Instruments for Higher Education." *Educational Assessment* 13 (1):56-75. doi: 10.1080/10627190801968240

Davis, L. E., Martin C. Harrison, A. S. Palipana, and J. P. Ward. 2005. "Assessment-driven learning of mathematics for engineering students." *International Journal of Electrical Engineering Education* 42 (1):63-72. doi: 10.7227/IJEEE.42.1.8

Delen, Erhan. 2015. "Enhancing a Computer-Based Environment with Optimum Item Response Time." *Eurasia Journal of Mathematics, Science and Technology Education* 11 (6):1457-1472. doi: 10.12973/eurasia.2015.1404a

DeVellis, Robert F. 2012. *Scale Development: Theory and Applications*. 3rd ed. London: Sage.

Dobson, Annette J. , and Adrian G. Barnett. 2008. *An Introduction to Generalized Linear Models*, *Texts in Statistical Science*. Boca Raton, FL: Chapman & Hall/CRC Press.

Hambleton, Ronald K, Hariharan H Swaminathan, and Jane Rogers. 1991. *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Harlen, Wynne. 2005. "Trusting teachers' judgement: research evidence of the reliability and validity of teachers' assessment used for summative purposes." *Research Papers in Education* 20 (3):245-270. doi: 10.1080/02671520500193744.

Hughes, Clair. 2013. "A case study of assessment of graduate learning outcomes at the programme, course and task level." *Assessment and Evaluation in Higher Eduction* 38:492-506. doi: 10.1080/02602938.2012.658020.

Irwin, Brian, and Stuart Hepplestone. 2012. "Examining increased flexibility in assessment formats." *Assessment & Evaluation in Higher Education* 37 (7):773-785. doi: 10.1080/02602938.2011.573842.

Jensen, Jamie L., Dane A. Berry, and Tyler A. Kummer. 2013. "Investigating the Effects of Exam Length on Performance and Cognitive Fatigue." *PLoS ONE* 8 (8):1-9. doi: 10.1371/journal.pone.0070270.

Kuo, Bor-Chen, Muslem Daud, and Chih-Wei Yang. 2015. "Multidimensional Computerized Adaptive Testing for Indonesia Junior High School Biology." *Eurasia Journal of Mathematics, Science and Technology Education* 11 (5):1105-1118. doi: 10.12973/eurasia.2015.1384a

Lord, Frederic M. 1952. *A Theory of Test Scores* Vol. 7, *Psychometric Monograph*. Richmond, VA.

Lord, Frederic M. 1953. "The relation of test score to the trait underlying the test." *Educational and Psychological Measurement* 13:517-548.

Lord, Frederic M. 1980. *Applications of Item Response Theory to Practical Testing Problems*. London: Routledge.

Lord, Frederic M., and Melvin R. Novick. 1968. *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Mumford, Christine L. 2010. "A multiobjective framework for heavily constrained examination timetabling problems." *Annals of Operations Research* 180 (1):3-31. doi: 10.1007/s10479-008-0490-3.

Muraki, E. . 1997. "A generalized partial credit model." In *Handbook of modern item response theory*, edited by W. van der Linden and R. K. Hambleton, 153-164. New York: Springer.

PARSCALE (Version 4.1). Scientific Software International, Lincolnwood, IK.

Nijmegen, Radboud University. 2011. "Conversion of Grades." Accessed 17 July 2014. http://www.ru.nl/io/english/general_0/document/.

Palmer, Edward J., Paul Duggan, Peter G. Devitt, and Rohan Russell. 2010. "The modified essay question: its exit from the exit examination?" *Medical Teacher* 32 (7):e300-e307. doi: 10.3109/0142159X.2010.488705.

Rue, H. 2014. "The R-INLA project." Accessed 17 July 2014. http://www.r-inla.org/.

Russell, Jill, Lewis Elton, Deborah Swinglehurst, and Trisha Greenhalgh. 2006. "Using the online environment in assessment for learning: a case-study of a web-based course in primary care." *Assessment & Evaluation in Higher Education* 31 (4):465-478. doi: 10.1080/02602930600679209.

Sadler, D. Royce. 2009. "Grade Integrity and the Representation of Academic Achievement." *Studies in Higher Education* 34 (7):807-826. doi: 10.1080/03075070802706553

Simpson, Lucy, and Jo-Anne Baird. 2013. "Perceptions of trust in public examinations." *Oxford Review of Education* 39 (1):17-35. doi: 10.1080/03054985.2012.760264.

Vu, Nv, A. Baroffio, P. Huber, C. Layat, M. Gerbase, and M. Nendaz. 2006. "Assessing clinical competence: a pilot project to evaluate the feasibility of a standardized patient -- based practical examination as a component of the Swiss certification process." *Swiss Medical Weekly* 136 (25-26):392-399.

Vukasovic, Martina. 2013. "Change of higher education in response to European pressures: conceptualization and operationalization of Europeanization of higher education." *Higher Education* 66 (3):311-324. doi: 10.1007/s10734-012-9606-4.

Westerheijden, Don F., Bjørn Stensaker, Maria J. Rosa, and Anne Corbett. 2014. "Next Generations, Catwalks, Random Walks and Arms Races: Conceptualising the development of quality assurance schemes." *European Journal of Education* 49 (3):421-434. doi: 10.1111/ejed.12071.

William, Dylan. 1996. "Standards in examinations: a matter of trust?" *The Curriculum Journal* 7 (3):293-306. doi: 10.1080/0958517960070303

Wittek, Line, and Tone Kvernbekk. 2011. "On the problems of asking for a definition of quality in education." *Scandinavian Journal of Educational Research* 55 (6):671-684. doi: 10.1080/00313831.2011.594618.

◈◇◈

APPENDIX

Beta Regression model

In this section, we provide a more detailed description of the beta regression model used in this paper. Let S denote a stochastic variable for the score on an arbitrary exercise for an arbitrary student, normalized to the [0, 1] interval. We assume that S is beta-distributed

$$\pi(s) = \frac{1}{B(a,b)} s^{a-1} (1-s)^{b-1}, \qquad a > 0, b > 0$$

where

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

where $\Gamma(x)$ is the Gamma function.

We now link the expectation of the beta distribution to a linear predictor of some covariates using a link function. We use the reparameterization

$$\mu = \frac{a}{a+b}, \quad 0 < \mu < 1$$
$$\phi = a + b, \quad \phi > 0$$

to arrive at

$$E(S) = \mu$$

Further, we get

$$Var(S) = \frac{\mu(1-\mu)}{1+\phi}$$

where ϕ is known as the precision parameter, since for a fixed μ, the larger ϕ, the smaller the variance in S. We link μ to a the linear predictor η using the logit-link function

$$\mu = \frac{e^{\eta}}{1+e^{\eta}}$$

Such a model is called beta regression. We use the following linear predictor in the beta regression model

$$\eta = k + M + D$$

where k is the fixed effect interception and M and D random effects representing the variability in the level of mastery (M) of students taking the exam and the difficulty level (D) of the exercises in the exam, respectively. We assume that M and D are normally distributed with zero expectations and variances $1/\tau_M$ and $1/\tau_D$, respectively. The parameters τ_M and τ_D are the inverse of the variance, which is referred to as precision. We assume that, given the random effects, the observations (score for a particular exercise for a particular student) are independent. We use a Bayesian approach and add prior distributions to the unknown parameters ϕ, k, τ_M and $\tau_{D.}$ For details about the prior distributions, we refer to the INLA web page (Rue, 2014).

Estimated values of parameters in a regression model

Table 2 shows properties of the posterior distributions of the variables ϕ, k, τ_M and τ_D for one of the 23 exams.

We see that k is less than zero, showing that, on average, the students scored below 0.5 on the exercises in this particular exam. We also observe that the largest estimation uncertainty is in the estimation of τ_D, variability in levels of difficulty on the exercises.

Table 2: Properties of the posterior distributions for the variables ϕ, k, τ_M and $\tau_{D.}$

Variable	Mean	Stdev	5% quantile	50% quantile	95% quantile
k	– 0.461	0.165	– 0.732	– 0.461	–0.191
ϕ	1.168	0.024	1.128	1.168	1.208
τ_M	1.357	0.135	1.148	1.349	1.592
τ_D	2.478	0.821	1.339	2.374	3.984