Construction of Psychometrically Sound Written University Exams

Andreas Frey[1,2], Christian Spoden[3] & Sebastian Born[4]

[1] Goethe University Frankfurt, Germany

[2] Centre for Educational Measurement (CEMO) at the University of Oslo, Norway

[3] German Institute for Adult Education - Leibniz Centre for Lifelong Learning, Bonn, Germany

[4] Center for Sepsis Control and Care - Jena University Hospital, Jena, Germany

Abstract:
Written university exams typically used at German-speaking universities often do not represent the learning objectives of the respective course appropriately. Moreover, they do not allow for criterion-referenced inferences regarding the degree to which the learning objectives have been met, and they are statistically unconnected across different test cycles. To overcome these shortcomings, we propose applying a combination of established methods from the fields of educational measurement and psychometrics to written university exams. The key elements of the proposed procedure are (a) the definition of the content domain of interest in relation to the learning objectives of the course, (b) the specification of an assessment framework, (c) the operationalization of the assessment framework with test items, (d) the standardized administration of the exam, (e) the scaling of gathered responses with item response theory models, and (f) the setting of grade levels with standard-setting procedures. Empirical results obtained from six test cycles of a real university exam at the end of an introductory course on research methods in education show that this procedure can successfully be applied in a typical university setting. It was possible to constitute a reliable and valid scale and maintain it across the six test cycles based on a common item nonequivalent group design. The comparison of the observed student competence distributions across the six years gave interesting insights that can be used to optimize the course.

Keywords:
item response theory, higher education, testing, measurement

Construction of Psychometrically Sound Written University Exams

Written university exams are broadly used to determine whether students have acquired the competences regarded as necessary to justify pass-decisions and to assign credit points. They are typically administered as more or less standardized tests with item types ranging from essays to multiple choice (MC) items to true-false items. Even though many written university exams resemble psychometric tests at first sight, unfortunately, they frequently do not meet common measurement standards in German-speaking countries. This might be due to a lack of awareness about the possible consequences of crude measurement habits and a lack of well-conceived alternative procedures for the construction of university exams. This situation is problematic because decisions with a high individual relevance for the students are often connected with the exam results.

We see three major problems with the typical written university exams in German-speaking countries. First, the learning objectives are often not systematically represented by the exam items with regard to cognitive demand and content. Thus, the extent to which the exam measures what it is supposed to measure, and therefore its validity (e.g., Hartig, Frey, & Jude, 2020), is unclear.

Second, the percentages of correct answers are often directly transformed into grades without reference to item content. This is inappropriate for university exams because grades should allow for interpretations about the degree to which students have mastered the competence-oriented learning objectives. Thus, unequivocal statements about what students know and can do should be possible. This can be accomplished by adopting criterion-referenced testing principles (Herzberg & Frey, 2011). Third, the reporting scales of written university exams are typically not statistically connected across the cohorts taking the test. This is reflected by the current statistical software that is available for the preparation of exams and the analyses of exam data (e.g., Muche, Janz, Einsiedler, & Mayer, 2013; Zeileis, Umlauf, & Leisch, 2014); this software does not include methods to connect scales across student cohorts. Without such a connection, exams are unfair because the same performance can lead to different grades in different applications of the exam if different items are used. This is a common problem university teachers face, when they have to discuss the results with students who argue that an exam was harder at a second testing point than at the first testing point. Without appropriate statistical procedures, these allegations cannot be adequately refuted. Certainly, there are examples of state-of-the art university exams that use appropriate psychometric principles, but the vast majority of university exams is subject to at least one or two of the above mentioned problems. These problems are not new. Similar and additional concerns have been raised, for example, by Atkins, Beattie, and Dockerell (1993) and by Elton (2004). Nonetheless, the issue is far from being resolved even though the areas of educational measurement and psychometrics provide solutions to the problems mentioned above.

With this article, we therefore want to (a) draw the attention of measurement experts and the staff involved in university teaching to this issue, (b) outline a procedure to overcome the mentioned shortcomings, and (c) illustrate this procedure with empirical results. With regard to (a), it should be noted that measurement experts, who are very familiar with the underlying concepts but unsure about whether they can be applied to university exams, and the staff involved in university teaching, who are in need of more guidance concerning the psychometric and measurement details, are confronted with different tasks when it comes to designing psychometrically sound exams. We aim to provide useful information for both groups. With regard to (b), the problem is not that appropriate methods are not available. One major aspect hindering the application of the available methods is that a directly applicable combination of compatible methods in the sense of a broadly applicable measurement standard for written university exams has not been formulated and empirically examined so far. In this article, we describe such a standard and its practical application. In order to keep the hurdle for an application of the proposed procedure low, we restrict our considerations to paper-based testing, which is currently still the predominant delivery format for written university exams (for a computer-based extension of the approach, see Frey, Spoden, Fink, & Born, 2020).

The text is organized as follows. First, the proposed procedure is described. Then, empirical results are presented that illustrate the application of the proposed procedure across six applications of a regular written university exam. The text ends with a brief discussion of the results and practical recommendations.

Proposed Procedure

To overcome the problems associated with typical written university exams in German-speaking countries, we propose combining some well-established methods and standards from educational measurement and psychometrics with the aim of ensuring that written university exams are sound measurements. To do this, specifications need to be made regarding the content domain of interest, the exam itself, the setting under which the exam is given, the psychometric scaling method, and the strategy used for assigning grades. Useful additional guidance can be found in research that describes structured approaches for the construction of competence tests (e.g., Frey & Hartig, 2019).

Regarding the test content, what exactly should be measured by the exam needs to be clearly specified. Thus, the *content domain of interest* has to be defined. A good point to start from is the learning objectives of the course. For university exams, the learning objectives are often that students should be able to successfully apply several cognitive processes in different parts of the content domain of interest. The parts nested in the content domain of interest are called content areas in the rest of this article. Ideally, these learning objectives are exemplified in terms of can-do statements in the description of the module or the course the exam is linked to. In order to organize the content domain of interest and to facilitate its communication to the students, it is useful to map it onto a matrix. Different content areas (e.g., chapters of a textbook) can be assigned, for example, to the lines of the matrix and different cognitive processes to the columns. One scheme that is widely employed for the categorization of cognitive processes in test construction is the taxonomy of Bloom (Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956). This taxonomy distinguishes between knowledge, comprehension, application, analysis, synthesis, and evaluation. Other schemes for structuring cognitive processes also exist; the one that corresponds best with the learning objective should be used. The resulting matrix is frequently called the assessment framework. If the different cells of the assessment framework are of varying importance for reaching the learning objectives, cell-specific weights can be assigned (Osterlind, 2002). Note that the assessment framework does not have to follow the described structure. It can take any form from a simple list to the described matrix to a multidimensional structure. The suggested two-dimensional matrix (cognitive processes × content areas) often provides a good balance between describing what has to be measured and being applicable at universities.

Next, the cells of the assessment framework are operationalized by tasks, referred to as *items* in the rest of this article. If the assessment framework is specified by a cognitive process × content area matrix as proposed, to solve one of these items, students have to apply the respective cognitive process in the specified content area. If all cells of the assessment framework are covered by such items, the learning objectives are operationalized exhaustingly without measuring aspects that are irrelevant for the learning objectives. To fully reach this goal, the item construction process has to be accurate and should include a review in order to generate items capable of assessing the behavior samples in the respective cognitive process × content area combination that they are supposed to operationalize. To keep the coding effort low and to facilitate to derive exam results from the assessment process, using items with closed response format wherever possible is advisable. Contrary to the opinion sometimes expressed, it is possible to construct closed format items that not only assess knowledge but also precisely measure higher-level cognitive processes. Previous research and handbooks on the construction of achievement tests give very useful recommendations on this (e.g., Haladyna & Rodriguez, 2013). Of course, there are also aspects that can only be measured with open items. Such items can also be used within the proposed procedure.

When it comes to the actual administration of the exam, a set of items representing the assigned weights should be presented to the students in a *standardized setting*. As is usual for university exams, answer copying should be prevented if the test is administered in group settings and only designated material must be used. Furthermore, the test items have to remain secret.

Thus, the students should not be allowed to take notes, take photos, or to keep the test booklets after the exam. In order to meet the wish sometimes expressed by students to gain insights into the test material beforehand, some sample items can be published some weeks before the exam.

The gathered responses then have to be *scored and scaled* to derive a pass-fail decision or to assign a grade. It is important to note that this step requires a more elaborated approach than is currently customary at German-speaking universities. The task at this point is to make inferences from the item responses about the extent to which the learning objectives have been met by the tested students. The derivation of criterion-oriented test score interpretations (Herzberg & Frey, 2011) is thus the goal here. These cannot be derived from simple sum scores, because items and persons are located on separate metrics. When using sum scores, it is possible to calculate the proportion of correct responses given to an individual item and the proportion of correct responses given by an individual student to the presented items. However, it is impossible to infer from these measures any kind of probability of an individual mastering certain kinds of tasks, problems, or exercises underlying the item's construction, even though this would be necessary for the desired criterion-referenced inferences. This issue can be resolved by using a measurement model capable of locating items and persons on the same common metric. In educational measurement and psychometrics, models from the item response theory (IRT; e.g., van der Linden, 2016) are typically used for such a purpose. A simple IRT model that is very useful for the scaling of data stemming from standardized university exams is the one-parameter logistic model (1PL). In the 1PL, the probability of a correct response u_{ij}=1 of person j=1,...,N with the latent ability level θ_j to an item i with difficulty b_i is defined as:

$$P(u_{ij} = 1|\theta_j, b_i) = \frac{\exp(\theta_j - b_i)}{1 + \exp(\theta_j - b_i)}$$

The 1PL has several advantages for university exams compared to more complex IRT models with additional item parameters and/or more dimensions. One important advantage is that the item difficulty b_i can be estimated precisely, even for relatively small samples. Estimates of additional item parameters, which are specified in more complex IRT models (e.g., item discrimination or pseudo-guessing parameter), tend to be less precise and—as a consequence— less stable across different test cycles and they also require remarkably larger sample sizes.

Unstable item parameter estimates would complicate or even preclude the connection of university exams across test cycles. However, in order to use the same standards for every cohort of students, such a connection needs to be established by using linking or equating strategies (e.g., Kolen & Brennan, 2014). Another advantage of the 1PL model lies in the fact that the sum of correct answers is a sufficient statistic in the 1PL model. Thus, the number of solved items can be reported to the students, which is a comprehensible scoring strategy. If partial solutions should be scored (e.g., usually useful when using constructed response items), the partial credit model (PCM; Masters, 1982) is an alternative to the 1PL. Its sample size requirements are also moderate, it usually provides stable parameter estimates as a prerequisite for stable linking across test cycles and it also allows score reporting at the sum-score level. Because the PCM is a generalization of the 1PL, it is also easy to switch to the 1PL if scaling problems occur. Recent developments in parameter estimation techniques such as Bayesian hierarchical modeling also make it possible to calibrate items even in very small samples (König, Spoden, & Frey, 2020). If the courses are too small to calibrate items even with these estimation techniques, there is still the option to wait one or two years, aggregate the responses across test cycles and then estimate item parameters from these data with a concurrent scaling approach. When some of these calibrated items are used in later exams, linking between cycles will be possible.

To make criterion-referenced test score interpretations possible, a *standard-setting procedure* (Cizek & Bunch, 2007, brief overview in Cizek, Bunch, & Koons, 2004, and some recent advances and applications in Blömeke & Gustafsson, 2017) can be used to define cut scores between grade levels. By defining cut scores, the degree to which the learning objectives are reached is directly placed on the IRT scale. Because the students are also located on this metric, their individual performance can easily be described by probabilities to solve certain items and, thereby, evaluates how well the learning objectives are met. A standard-setting procedure that is well suited for university exams is the bookmark procedure (e.g., Mitzel, Lewis, Patz, & Green, 2001). This can be applied in a simplified version with a panel that consists of only a few persons who are involved in the connected teaching (i.e., university staff) and that defines a limited number of cut scores (see Frey, Spoden, Born, & Fink, 2017).

Empirical Application

The outlined procedure was applied to the written paper-based exam for a lecture on "Introduction to Research Methods in Education" at the Friedrich Schiller University Jena, Germany, for six connected test cycles. In the following section, we will (a) illustrate the implementation of the concept in this real university setting, (b) analyze to what extent the important link from the first to the second test cycle worked, (c) demonstrate that results with scales linked across test cycles can provide new insights, and (d) discuss the conditions under which people with no training in psychometrics can use the procedure.

Method

Assessment Framework.

To specify the content domain of interest, the person responsible for the content of the course established the assessment framework. The learning objectives of the course were for students to be able to apply different cognitive processes in nine content areas.

The content areas covered fundamental aspects of quantitative and qualitative research methods in education. The cognitive processes used in the assessment framework were based on the taxonomy of Bloom, distinguishing between knowledge, comprehension, application, analysis, synthesis, and evaluation. The cognitive processes of knowledge, comprehension, and application were regarded as being more important for this introductory course (typically taken in the first semester) than the cognitive processes of analysis, synthesis, and evaluation. The latter processes are addressed later in the study program when basic knowledge has been acquired. It was therefore decided to develop items primarily for the first group of cognitive processes (116 items) compared to the latter cognitive processes (8 items). For other courses with other learning objectives, other cognitive processes might be more important.

Item Writing and Reviewing

The cells of the assessment framework were operationalized by test items. Staff members from the department of Research Methods in Education at the Friedrich Schiller University Jena including the person responsible for the content and the lecture constituted the item writing and reviewing team.

This team constructed contextualized items comparable to those used in international large-scale assessments of student achievement such as the Programme for International Student Assessment (PISA) or the Trends in International Mathematics and Science Study (TIMSS). The context of these items was a graphical object such as a photo, a graph, a chart, an illustration, a reading passage, a table, or combinations of these.

Each of the items was assigned to exactly one content area and to exactly one cognitive process. A larger batch of 49 items was constructed for the first test cycle in 2012, which was supplemented by smaller numbers of items in each of the following years. The item pool in 2017 included 124 items and is shown in Table 1.

The item development process for the first test cycle started with a brainstorming session in order to get ideas for interesting contexts and tasks with high relevance for educational practice and research. Next, reasonable stimuli and intelligible response options in closed or short format were formulated. The closed items were designed in multiple choice (MC) with only one correct answer and three distractors or complex multiple choice (CMC) formats (consisting of four to six questions which have to be answered with yes or no which all have to be answered correctly to be scored as correct). The items in the short response format required students to insert a brief answer such as a number, one or several words, or a formal expression of a statistical hypothesis. In order to keep the coding effort low, items in short response format were only constructed if the targeted cell of the assessment framework could not be operationalized by an item in closed format.

An iterative item review followed the item construction process. In this step, the items were reviewed by at least one team member other than the author of the item. Brief written feedback was provided on the item draft and used by the author for a revision of the item, if necessary. The process was iterated until all the concerns raised by the reviewer were resolved. The item writing and reviewing process in the first test cycle was completed in a meeting of the item writing and reviewing team. In this meeting, any remaining problems with the draft items were discussed and resolved and the set of items to be used in the exam was selected. In the following test cycles, the item construction and reviewing was carried out without additional brainstorming and final meetings.

Test Material

For the actual testing of the students, test booklets were assembled. Besides the test items, each booklet contained a cover page, instructions, a list of mathematical formula, and statistical tables for different distributions. In order to hinder answer copying and to roughly balance potential item position effects (e.g., Nagy, Nagengast, Frey, Becker, & Rose, 2019; Trendtel & Robitzsch, 2020), two different booklets were assembled for each test cycle. The test booklets contained 35 to 40 items. The sequence of the items in the booklets was chosen in such a way that the median item position of all items was comparable.

Test Setting and Procedure

The actual testing took place in a group setting at the end of the courses in February in each of the relevant years (2012–2017). The students were assigned randomly to seats in a large lecture hall. In order to prevent answer copying, there was at least one spare seat between each two students. Addition-

Table 1 Number of Items per Cell in the Assessment Framework

Content Area	Cognitive Process						
	Knowledge	Compre-hension	Application	Analysis	Synthesis	Evaluation	Sum
1. Basics in Research Methods	11	4	7	3	-	-	25
2. Quantitative Methods: Design	2	8	3	-	-	-	13
3. Quantitative Methods: Data Collection Methods	4	3	3	1	-	-	11
4. Qualitative Methods: Design	4	3	-	1	-	-	8
5. Qualitative Methods: Observation Methods	6	2	-	-	-	1	9
6. Qualitative Methods: Analysis	3	2	-	-	-	-	5
7. Descriptive Statistics	8	1	8	-	1	-	18
8. Inferential Statistics I: Probability and Distribution	3	1	11	-	-	-	15
9. Inferential Statistics II: Testing of Hypotheses	2	9	8	-	-	1	20
Sum	**43**	**33**	**40**	**5**	**1**	**2**	**124**

ally, the two paper-based booklets used per test cycle were alternately distributed to the participants, along with accessories for working on the test (two pencils, a ruler, and a calculator without a programming function). After the students had taken their seats, they received brief oral instructions and were given the possibility to ask questions regarding the exam procedure. After that, they had 80 minutes to work on the items and a further 10 minutes to transfer their solutions to the prepared answer sheet. Thus, the testing time per student was 90 minutes. The responses on the answer sheets were recorded automatically using the software TeleFORM®. The resulting data matrix was analyzed with the statistical packages SPSS for data management and general statistical analyses and ConQuest 3.01 for IRT scaling in the years 2012 to 2015. From 2016 on, all analyses were carried out in R.

Scaling and Linking

The exam was given to 84 ≤ *N* ≤ 129 students studying educational science as their major in the years 2012 to 2017. Some details on the samples are given in Table 2. In all years, the between 79% and 87% of the students were female. All students finished the exam and handed in their completed answer sheet.

The responses collected in the individual test cycles were scaled separately with the PCM. The item fit was evaluated by the mean squared error, the weighted mean squared error, and their corresponding *t*-values. The ability of the students was estimated with the weighted likelihood estimator (WLE; Magis & Verhelst, 2017; Warm, 1989). The ability estimates were transformed to 11 grade levels (1.0, 1.3, 1.7, 2.0, 2.3, 2.7, 3.0, 3.3, 3.7, 4.0, 5.0) using the cut scores derived from the stan-

Table 2 Sample Statistics for the University Exams From 2012–2017

Test	Student		Item		*θ*-distribution		Proportion
Cycle	Cohort	*N*	Number	Reliability	*M*	*SD*	Female
1	2012	114	37	.827	0.000	0.835	84%
2	2013	97	35	.879	-0.139	0.994	83%
3	2014	129	37	.809	-0.387	0.722	82%
4	2015	112	35	.781	0.067	0.722	87%
5	2016	103	40	.757	-0.070	0.663	79%
6	2017	84	38	.637	-0.222	0.475	82%

dard-setting procedure described in the next section. These grade levels are compulsory for graded written exams at the Friedrich Schiller University Jena. The best grade was a 1.0 and the worst passing grade was a 4.0. A grade of 5.0 indicated a fail.

With the aim of linking two adjacent test cycles, a common item nonequivalent group design (Kolen & Brennan, 2014) was used. For this purpose, 15 to 20 items used in the former test cycle were selected as link items. These items were selected to represent a broad range of the content specified by the assessment framework, to cover a broad difficulty range, and to have high point-biserial correlations with the total score of the test. When assembling the booklets for a subsequent test cycle, it was further ensured that the link items were presented in approximately the same position as in the previous test cycle. The linking procedure comprised four major steps. First, the responses to all items used in the second test cycle were scaled with the PCM and the item fit was evaluated. Second, a mean-mean equating (Kolen & Brennan, 2014) for the set of link items was conducted to transform the item difficulty estimates of the second test cycle to the IRT metric of the first test cycle. Third, link items showing significantly different difficulties in the two test cycles (i.e., item drift) were identified. It was then checked whether the 95% confidence interval around the difficulty of a link item estimated for the second test cycle covered the difficulty parameter of the same item from the first test cycle. If the confidence interval did not cover the item parameter estimate, the null hypothesis of item parameter invariance was rejected and the respective item was not regarded as a link item for the second test cycle. Fourth, the PCM was used to conduct a final scaling of the responses from the second test cycle with the difficulty parameters of the remaining invariant link items anchored at their values from the first test cycle. The difficulties of the other items were freely estimated. This kind of linking was carried out in PISA until 2012 (e.g., OECD, 2012).

Standard Setting

In order to determine cut scores between the grades, a simplified bookmark procedure was used, based on the scaling results from the first test cycle. Therefore, all presented items were ordered according to their difficulty estimate from the easiest to the most difficult item. This ordered booklet was then analyzed by a panel of three content experts who had been involved in teaching the subject area covered by the exam (university staff members).

Note that, for a typical standard-setting procedure with the bookmarking method, larger panels of 18 to 24 persons are recommended (Lewis, Mitzel, Green, & Patz, 1999). However, because, in the case of written university exams, a consensus only needs to be reached between the persons responsible for the curriculum of the course, for the learning objectives, and for the actual teaching—and not between the many stakeholders that are typically involved in national or international standard-based assessments—a relatively small panel can be used. The panelists were asked to identify two cut scores. The first was the cut score between *pass* and *fail*. To set this cut score, they started from the easiest item and determined the one item in the ordered item booklet that marked the threshold between pass and fail. Following the recommendation of Frey et al. (2017), the response probability (see OECD, 2012) was set to .80. Thus, the panelists had to agree that solving the item that marked the pass-fail threshold and all items before it (with a lower difficulty) with a probability of at least .80 would indicate sufficient fulfillment of the learning objectives. The difficulty estimate of the appointed item was used as the cut score between the grades 5.0 (fail) and 4.0 (pass). Second, the panelists were asked to set the cut score between the best grade (1.0) and the second best grade (1.3). In order to determine this cut score, they started from the most difficult item downwards in order to identify an agreed set of items that need to be solved with a probability of at least .80 to fully meet the learning objectives. The difficulty estimate of the easiest item of this collection was used as the cut score between the grades 1.0 and 1.3. The eight cut scores between the remaining grades were set at equidistant distances between the two extreme cut scores. Using equidistant distances was regarded as the most appropriate solution because no criterion-referenced assumptions could be made about the differences between the relatively fine-grained grades.

Results

First and Second Test Cycles

For the first test cycle, no item showed a significant misfit. One item was excluded because there were only incorrect answers, which rendered the further scaling of this item impossible. Thus, 36 items remained in the test for the following analyses.

The estimated mean of the difficulty distribution was -1.03 (*SD* = 1.50). The average point-biserial correlation between the single items and the sum of solved items (item discrimination from classical test theory) was .37 with a range of .09 to .61. The mean of the latent ability distribution was fixed to 0.00 for model identification purposes; the variance was estimated as 0.75.

In the second test cycle, one item showed a significant misfit (WMNSQ = 1.22; t = 2.6). However, because the item had an acceptable point-biserial correlation of .21 with the total test score (and because providing feedback to the students was easier without deleted items), it was kept in the test. The estimated mean of the second cycle's difficulty distribution was -0.69 (*SD* = 1.26). With a mean of .43 (range: .08 – .66), the point-biserial correlation between the single items and the total score was slightly higher than for the first test cycle. The mean of the latent ability distribution after linking was slightly lower (-0.11) compared to the first test cycle, whereas the variance (1.11) of the latent ability distribution was higher. With values of .80 (2012) and .85 (2013), the reliabilities of the ability estimates were good.

Concerning the linking between the two test cycles, for 12 of the 17 link items, the null hypothesis of item parameter invariance held ($p \geq .01$). The other five link items showed significantly different difficulties in the two test cycles. Their difficulty parameters were therefore not anchored to the values from the first test cycle and were instead estimated freely for the second test cycle.

Test Cycles Three to Six

The results in the following test cycles regarding item fit, linking stability, and reliability were comparable to the results obtained in the first two test cycles. A closer look at the distribution of the student abilities measured provides an interesting insight into the development across test cycles. Figure 1 shows the competence distributions of all six test cycles. Note that all test cycles are connected with a stable link so that the distributions in Figure 1 can be directly compared with each other. Obviously, the average competence and the variance was substa ntially lower in 2014 compared to the two years before. In 2015, the distribution strongly resembles the distributions from 2012 and 2013. In 2017, there is a noteworthy drop in the average competence.

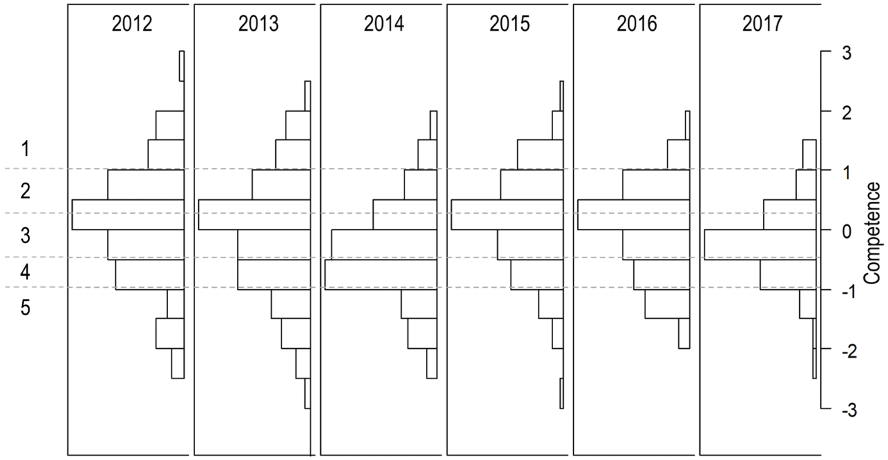

Figure 1 Student Competence Distributions Obtained by an IRT Model for Six Linked Standardized University Exams. The Numbers at the Left are the Grade Levels, with 1 being the Best and 5 the Worst Grade.

Discussion

In reaction to the claim that written university exams often do not meet common measurement standards, this article proposes a procedure to overcome this problem and illustrates this procedure with an empirical application. The suggested procedure combines established methods from educational measurement and psychometrics. The main features of the procedure are (a) the definition of the content domain of interest in relation to the learning objectives of the course, (b) the specification of an assessment framework, (c) the operationalization of the assessment framework with test items, (d) the standardized administration of the exam, (e) the scaling of the responses with an IRT model in the free software package R and with well-documented tutorials for IRT modeling (e.g., Chalmers, 2012; Robitzsch, Kiefer, & Wu, 2020), and (f) the setting of grade levels with standard-setting procedures. Thus, we advocate a rational procedure that makes it possible to directly connect test scores and/or grades with the extent to which the learning objectives of the respective course have been fulfilled. Furthermore, the procedure offers the possibility to keep the requirements of what students should know and can do to reach a certain grade level constant across test cycles.

Our empirical results show that the procedure can be successfully applied in a typical university setting. With values of around .80, the achieved reliabilities were good. Nevertheless, it has to be noted that the standard errors associated with the ability estimates were rather large. The average standard error amounted to about 0.40 and was thus approximately one and a half times as high as the section covered by one of the 11 grade levels (0.25). Thus, the 95% confidence interval around the ability estimate of a student typically covered several grade levels and the student was just placed in the most likely category. Therefore, it might be more appropriate to use a smaller number of grade levels in order to achieve precision in written exams of reasonable lengths. Large numbers of grade levels create the impression of high precision but—as can be seen by the standard errors in relation to the width of the grade intervals—this is not actually the case. An alternative option to increase the precision of the ability estimates is to use computerized adaptive testing (CAT; Frey, 2020) for item selection. In Germany, however, the use of CAT for exams is still a gray area (see Frey et al., 2020 for a discussion), because no corresponding court decision has yet been made. However, as first approaches for using CAT in university exams already exist (Spoden, Frey, Fink, & Naumann, 2020), this should soon be clear.

This study shows that it is possible to link written university exams successfully across many years. Changes were made to the course every year. There were new examples and new exercises, but the basic cognitive processes × content areas structure always remained the same. This orientation resulted in a stable measurement instrument and made it possible to learn more about the determinants of the learning success of the students. The dip in average competence observed in 2014, for example, can be traced back to a reduction in the number of tutorials that accompanied the lecture from two to one. After two tutorials were again offered in 2015, the average competence increased again, up to the value observed before 2014. The remarkably low competences found in 2017 coincide with the change in the content taught in a course in a minor subject. About half of the students who took the exam analyzed in this paper attended this course. Previously in this course, overlapping course contents were taught, which were then replaced in 2017 by other content so that the learning opportunities relevant for the analyzed exam became smaller. Even though causal inferences cannot be drawn based on the available data, the linked scales offer good opportunities to monitor the development of the student competences across test cycles and to formulate assumptions and hypotheses.

With this article, we propose one applicable solution for how to reach an appropriate measurement standard for written university exams. Lecturers (and students) benefit in several ways from the new procedure: The proposed procedure makes it possible to draw criterion-referenced inferences and thereby to fulfill the requirements of competence-oriented exams specified in the Bologna Process and recently emphasized by the German (Universities) Rectors' Conference (Hochschulrektorenkonferenz, 2015). The procedure also makes it possible to use the same learning objective-related evaluation criteria across test cycles. Thus, the assigned grades are independent of the achievement level of the tested cohort. This means an increase in the exam's fairness, which, in turn, is a prerequisite for the validity of the interpretations and uses derived from its results. The increase in fairness can be illustrated to the students; this is likely to foster a more positive perception of exams in general within the group of students and it also makes a strong case for possible appeals against grading decisions. Given that waiting for an easier exam in coming years becomes pointless, this procedure might even help to reduce students' procrastination in the long run.

As a concluding remark, it is worth noting that measurement experts and psychometricians so far have not directed enough attention to the habits of university staff in preparing and constructing university exams. Although this article exemplifies a directly applicable combination of methods in the sense of a broadly applicable measurement standard for written university exams, more research is certainly needed to analyze these habits and investigate the personal and situational conditions necessary for the construction of psychometrically sound exams.

References

Atkins, M. J., Beattie, J., & Dockerell, W. B. (1993). *Assessment issues in higher education*. London: Department of Employment.

Blömeke, S., & Gustafsson, J.-E. (2017). *Standard setting in education. The Nordic countries in an international perspective.* Springer.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain*. New York, Toronto: Longmans, Green.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48(6),* 1–29.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks: Sage.

Cizek, G. J., Bunch, M. B., & Koons, H. (2004). Setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice*, *23*, 31–50.

Elton, L. (2004). A challenge to established assessment practice. *Higher Education Quarterly*, *58*, 43–62.

Frey A. (2020). Computerisiertes adaptives Testen [Computerized adaptive testing]. In H. Moosbrugger & A. Kelava (Eds.), *Testtheorie und Fragebogenkonstruktion* (3rd Ed., pp. 501–524). Springer. https://doi.org/10.1007/978-3-662-61532-4_20

Frey, A., & Hartig, J. (2019). Kompetenzdiagnostik [Competence Measurement]. In M. Harring, M., C. Rohlfs & M. Gläser-Zikuda (Eds.), *Handbuch Schulpädagogik* (pp. 849–858). Münster: Waxmann.

Frey, A., Spoden, C., Born, S., & Fink, A. (2017). Konstruktion psychometrisch fundierter Hochschulklausuren für das digitale 21. Jahrhundert [Construction of psychometrically sound university exams for the digital 21st century]. Jena: Friedrich Schiller University.

Frey, A., Spoden, C., Fink, A., & Born, S. (2020). Kompetenzorientierte individualisierte Hochschulklausuren und deren prüfungsrechtliche Einordnung [Competence-referenced individualized written exams in higher education and their classification under German examination law]. *eleed*, *13*. Retrievable from: urn:nbn:de:0009 5-51197.

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating multiple-choice test items*. New York: Taylor & Francis.

Hartig, J., Frey, A., & Jude, N. (2020). Validität von Testwertinterpretationen [Validity of test score interpretations]. In H. Moosbrugger & A. Kelava (Eds.), *Testtheorie und Fragebogenkonstruktion* (3rd Ed., pp. 529–545). Springer. https://doi.org/10.1007/978-3-662-61532-4_21

Herzberg, P. Y., & Frey, A. (2011). Kriteriumsorientierte Diagnostik. In L. F. Hornke, M. Amelang & M. Kersting (Eds.), *Methoden der psychologischen Diagnostik. Enzyklopädie der Psychologie*, B/II/2 (pp. 281–324). Göttingen: Hogrefe.

Hochschulrektorenkonferenz (2015). nexus impulse für die Praxis Nr. 4: Kompetenzorientiert Prüfen. Zum Lernergebnis passende Prüfungsaufgaben [nexus impulses for practice nr. 4: competence-oriented teaching. Aligning learning results and exams]. Retrievable from: https://www.hrk-nexus.de/fileadmin/redaktion/hrk-nexus/07-Downloads/07-02-Publikationen/HRK_Ausgabe_4_Internet.pdf

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking. Methods and practices* (3rd Ed.). New York: Springer.

König, C., Spoden, C., & Frey, A. (2020). An optimized Bayesian hierarchical two-parameter logistic model for small-sample item calibration. *Applied Psychological Measurement*, *44*, 311–326.

Lewis, D. M., Mitzel, H. C., Green, D. R., & Patz, R. J. (1999). *The bookmark standard setting procedure*. Monterey: McGraw-Hill.

Magis, D., & Verhelst, N. (2017). On the finiteness of the weighted likelihood estimator of ability. *Psychometrika, 82***,** 637–647.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249–281). Mahwah: Erlbaum.

Muche, R., Janz, B., Einsiedler, B., & Mayer, B. (2013). Ein (halb-)automatisiertes Prüfungstool für semesterbegleitende Prüfungen im Fach Biometrie (Q1) im Medizinstudium [A (semi-)automatic tool for semester accompanying exams in the biometry course in medical studies]. *GMS Medizinische Informatik, Biometrie und Epidemiologie*, *9*(3), Doc11 (20130517).

Nagy, G., Nagengast, B., Frey, A., Becker, M., & Rose, N. (2019). A multilevel study of position effects in PISA achievement tests: Student- and school-level predictors in the German tracked school system. *Assessment in Education: Principles, Policy & Practice*, *26*, 422–443.

OECD (2012). *PISA 2009 technical report*. Paris: Author.

Osterlind, S. J. (2002). *Constructing Test Items: Multiple-choice, constructed-response, performance, and other formats* (2nd Ed.). Boston, Dordrecht, London: Kluwer.

Robitzsch, A., Kiefer, T., & Wu, M. (2020). *TAM: Test Analysis Modules*. R package version 3.5-19, https://CRAN.R-project.org/package=TAM.

Spoden, C., Frey, A., Fink, A., & Naumann, P. (2020). Kompetenzorientierte elektronische Hochschulklausuren im Studium des Lehramts [Competence-related electronic written exams in higher education teacher training programs]. In K. Kaspar, M. Becker-Mrotzek, S. Hofhues, J. König, & D. Schmeinck (Eds.), *Bildung, Schule und Digitalisierung* (pp. 184–189). Münster: Waxmann.

Trendtel, M., & Robitzsch, A. (2020). A Bayesian item response model for examining item position effects in complex survey data. *Journal of Educational and Behavioral Statistics*. Advance Online Publication.

van der Linden, W. J. (Ed.). (2016). *Handbook of item response theory. Volume one: Models*. Boca Raton: Chapman & Hall/CRC.

Warm T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*, 427–450.

Zeileis, A., Umlauf, N., & Leisch, F. (2014). Flexible Generation of E-Learning Exams in R: Moodle Quizzes, OLAT Assessments, and Beyond. *Journal of Statistical Software*, *58*(1), 1–36.

The preparation of this article was supported in part by grants 01PK11005C (MoKoMasch) and 16DHL1005 (KAT-HS) from the German Federal Ministry of Education and Research.

On behalf of all authors, the corresponding author states that there is no conflict of interest.

Corresponding author:
Andreas Frey, PhD
Goethe University Frankfurt
Theodor-W.-Adorno-Platz 6
60323 Frankfurt
Germany
frey@psych.uni-frankfurt.de