Measures for Quality Assurance of Electronic Examinations in a Veterinary Medical Curriculum

Robin Richter ■ Andrea Tipold ■ Elisabeth Schaper

ABSTRACT

Since 2008, electronic examinations have been conducted at the University of Veterinary Medicine Hannover, Germany which are analyzed extensively in the current study. The aim is to assess the quality of examinations, the status quo of the electronic examination system and the implementation of recommendations regarding the conduct of exams at the TiHo. Based on the results suitable indicators for the evaluation of examinations and items as well as adequate quality assurance measures and item formats are to be identified. For this purpose, 294 electronic examinations carried out from 2008 to 2022 of the veterinary medicine course with an average of 248 participants each were evaluated with regard to the quality criteria reliability, difficulty index, and discrimination index. The main finding was that the number of items and the proportion of reused questions were identified as factors through which the quality of the examinations can be increased with simple adjustments. A higher number of items led to better reliability, whereby the required minimum reliability in examinations of 0.8 was reliably achieved from an item number of 98 questions. The proportion of reused questions should be kept low, as these had a negative influence on the characteristic values. Measures accompanying examinations, such as training of question authors and a pre- and post-review process, should also ensure the quality of examinations. For the post-review process, the distribution of examination results, reliability, item and distractor analysis are adequate indicators for evaluating examinations.

Key words: e-assessment, veterinary education, examinations, item formats, Cronbach's a, discrimination index, multiple choice questions (MCQ)

INTRODUCTION

Summative examinations are used for decision-making to determine whether defined objectives of a learning section or training phase have been achieved. The students can be admitted to further training stages or finally obtain a degree, which in the university environment is realized, among other things, via graded performance assessments that decide on "pass" or "fail".[1,2] Especially in the health care sector, such examinations bear a high responsibility, since they are intended to ensure, on the one hand, that the candidates possess the knowledge and skills to be able to practice their profession safely[3,4] and, on the other hand, failure has serious consequences for the further training and professional career.[2] Accordingly, such examinations should meet high standards with regard to the three quality characteristics reliability, objectivity and validity.[5–11]

Reliability describes the reproducibility of test results, which means that, in theory, candidates should obtain the same results under the same circumstances when they take a well-designed test again.[9,11,12] In the context of testing, reproducibility is considered to play an important role, as it is essential for the interpretation of test results, the decision on whether to pass, and the credibility of the results.[8,13,14] As a quality criterion, reliability can be measured by various methods.[8,14,15] The most frequently used method for quantifying reliability is the variance-analytical method developed by Cronbach[16] for determining internal consistency, which is called Cronbach's α.[17] The variable α takes a value between 0 and 1, with a higher value describing better reliability; usually a minimum reliability of 0.8 to 0.9 is required for summative tests,[2,6,10,13,18,19] in some cases ranges of 0.7 and higher are also deemed as sufficient.[13,17]

Objectivity characterizes the independence of the results of the examination from the examiners and the framework conditions of the examination,[2] which is usually achieved in written examinations by the fact that there is an answer for each question that is unanimously regarded as correct by representatives of the specialist subject. At the same time, this lack of leeway for subjective assessments and interpretations means that the examination results can be automatically evaluated by machines.[12]

Validity refers to the ability of a test procedure to measure exactly the construct it is designed to measure.[14,20,21] In the example of the examination system, this term refers to the extent to which an examination can suitably test the defined objectives of a training section. Furthermore, validity is further differentiated[14,21]:

- Content: The test must be representative of the entire area of knowledge to be covered.
- Prognostic: Performance on this test allows conclusions to be drawn about the candidate's future performance.
- Concurrent: The ability of the test to evaluate the examinee's current performance is assessed by comparing it to a previously validated test.
- Construct validity: The result of the test correlates with the knowledge and ability level of the candidates.

In this respect, the potential of electronic examinations can be used to improve the quality, effectiveness, and efficacy of examination procedures[9,22–27] and, consequently, the quality criteria. At the same time, in addition to the possible positive impact on quality, special requirements also arise with regard to the didactics, methodology, and organization of examinations.[24]

An important tool for improving the quality, effectiveness, and efficacy of examination procedures as well as the organization of exams are the automation processes of electronic exam management platforms, including item analysis of exam questions with respect to the two parameters difficulty index and discrimination index. **Difficulty** is defined as the relative frequency with which candidates choose the correct answer[28] and is thus expressed as the percentage of candidates with correct answers to the exam question.[2] International literature mentions varying optimal ranges for item difficulty, for example 40%–80%,[19] 41%–94%[2] as well as 30%–70%.[29] Krebs[2] states that items with a discrimination index that show a significant positive contribution to performance differentiation occur predominantly in the range of a difficulty of 42%–93% and a missing to negative contribution to differentiation is observed below 40% and above 95%, respectively. Hermi and Achour,[29] however, limit the recommendation to a narrower range of difficulty where the best values of discrimination index occur. Since the purpose of scattering the item difficulties in these recommended optimal ranges is to generate items with good performance differentiation properties it is not a mandatory requirement for items to fall within these defined segments. As most examinations in the veterinary medical curriculum are criterion-referenced tests it is also necessary and reasonable to examine core knowledge and competencies, which in turn might lead to relatively easy items with a low discrimination index.[19] Consequently, exam coordinators should strive for a balance of easy and difficult items. **Discrimination index** characterizes the ability of an item to distinguish between candidates with good performance and those with poorer performance; accordingly, an item with a good discrimination index is answered correctly by good candidates and incorrectly by poorer candidates.[21] The discrimination index scale ranges from −1 to +1 and is divided into categories >0.4 "Very good", 0.3–0.39 "Good", 0.2–0.29 "Acceptable", and ≤0.19 "Poor",[30] while in more recent literature the range of 0.1–0.19 is categorized as "Critical"[19] or "Weak discrimination index"[2] and only values from <0.1 are evaluated as "Poor".[19]

In 2008, the first electronic examinations (e-examinations) in the state examination were introduced at the University of Veterinary Medicine Hannover, Foundation (TiHo) using the Q [kjuː] system from the external provider Codiplan GmbH, Bergisch Gladbach, Germany.[31] In addition to a possible improvement in efficacy, the focus of the deployment was primarily on an increase in quality, taking into account the quality criteria and expansion potential of the platform through new examination and question formats.[32]

With the aim of further developing and optimizing the e-assessment as well as the measures for quality assurance (QA) of e-exams, a system changeover to Q-Exam Institution of the same provider (now known as IQuL GmbH, Bergisch Gladbach) was carried out in 2017, taking into account the recommendations of Jünger and Just,[6] whereby, among other things, a multi-stage review process as well as blueprints for all individual exams were introduced to ensure the validity, quality, and comparability of the exams.[5,6,33]

The aim of this study is to record and evaluate the quality of the examinations using Classical Test Theory (CTT), the status quo of the electronic examination system and the implementation of the recommendations at the TiHo. For this purpose, the following hypothesis was tested: the majority of electronic examinations at the TiHo follow the recommended standards for quality characteristics.[2,6,19] Based on these data, further development and optimization of the examination process will be undertaken, and suitable indicators for evaluating examinations and items for everyday practice will be identified, whereby adjustments can be implemented through simple changes.

MATERIAL AND METHODS

The examination management platform Q-Exam Institution acts as the examination and question database; the software Q-Examiner (IQuL GmbH, Bergisch Gladbach) is used to conduct the electronic examinations at the TiHo.

For evaluating the **Cronbach's α**, all examinations of the study program of veterinary medicine at the TiHo conducted electronically since 2008 until mid-2022 with a number of participants of more than 100 students each were selected. A total of 294 data sets were evaluated in the form of raw data for statistical item analysis of 31 examinations conducted annually with varying question compositions from a total of 28 subject areas with an average of 248 participants (ranging from 101 to 306 participants).

Furthermore, all examinations conducted between 2008 and mid-2022 that can be assigned to the state examination as well as the first and second preclinical examination were used for the analysis with regard to **question formats and their characteristic values** (see Figure 1). Examinations from 28 subject areas with an average number of participants of 248 (ranging from 101 to 306) were available.

The five item formats primarily used at the TiHo, Multiple Choice Question (MCQ) Type A, Kprim,[2] Key Feature, Picture diagnosis, and Picture mapping, were considered.

For a more detailed examination of the **Kprim** item format, a sample of 22 data sets of exams from five subject areas with a total of 332 Kprim items was selected, since Kprim items are established in these exams.

Finally, to investigate the number and effects of **reused items**, all electronic exams from the state exam, the first preclinical examination and second preclinical examination from 2017 to mid-2022 were examined, as the Q-Exam Institution system makes it easier to detect item re-use. Thus, a total of 133 exams from the new system from 28 subject areas with an average of 252 participants (ranging from 101 to 304) and a total of 7,837 items were the focus of this study.

Descriptive analysis was performed using the Microsoft Office Excel 2010 spreadsheet program (Microsoft Corporation, Redmond, WA, USA) and advanced statistics were performed using SAS software, version 9.4 and the SAS Enterprise Guide 7.1 (SAS Institute Inc., Cary, NC, USA).

For further analysis of test reliabilities using SAS software, Cronbach's α values were checked for normal distribution. Here, the significance level for all tests was 5%, meaning that a p value <.05 was interpreted as a significant result. The Shapiro–Wilk test for normal distribution was p=.147, so the null hypothesis must be accepted here, that is, the data were normally distributed. Based on this, a two-sample t-test was performed, with the classification variable time of testing (before or after system conversion) and the analysis variable

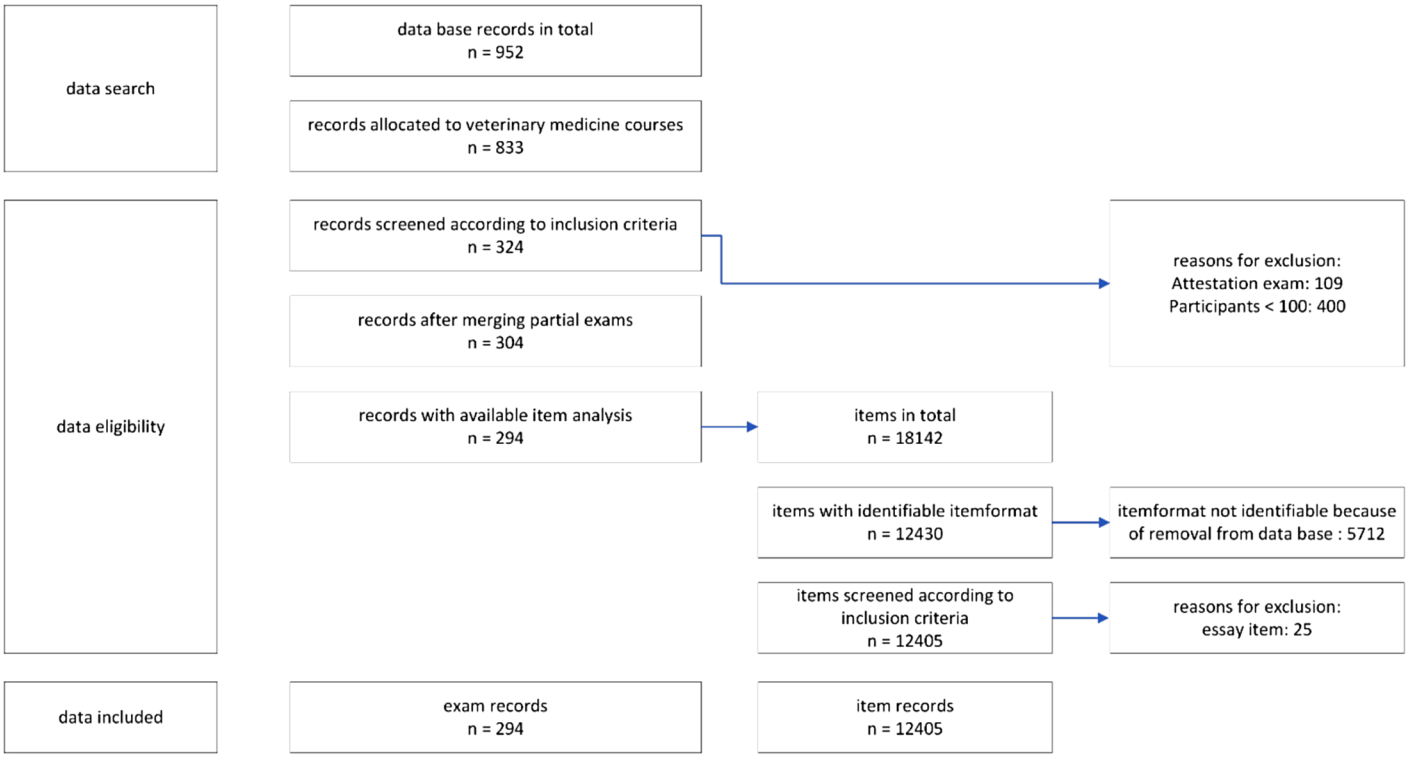

Figure 1: Data compilation and preparation for the evaluation of item formats and their characteristic values.

Table 1: Characteristics and scoring scheme of the five item formats used at the TiHo.

Question format	Description	Evaluation	Comment
MCQ Type A	A single-choice, best-answer format consisting of one correct answer (*attractor*) and two to four incorrect options (*distractors*)	One point for selecting an *attractor*. No points for selecting a *distractor*	Variable number of choice options, categorized by their number as follows: "MCQ Type A3", "MCQ Type A4", and "MCQ Type A5"
Kprim	A true-false selection procedure where "correct" or "incorrect" must be selected for each of four answer options	Four correct matches: one point. Three correct matches: half a point. Fewer than three correct matches: no points	Kprim is mainly used in German-speaking countries
Key feature	A group of three items that build on each other in terms of content or topic. The order is predetermined and after initial processing the selected choice option cannot be changed.	One point per correctly answered subquestion	
Picture diagnosis	A marker is placed on an image	One point if the marker is in the predefined area	
Picture mapping	Predefined terms are assigned to given pixels	One point if the assignment of the terms is completely correct. Half a point if at least 50% of the terms are correctly assigned	

reliability. To analyze the differences in mean discrimination index of the item formats, a Kruskal–Wallis test as well as Dwass–Steel–Critchlow–Fligner test was performed. Regarding the influence of the number of items in examinations and the discrimination index on items, Spearman's rank correlation was calculated. Furthermore, to assess the changes of item difficulty and item discrimination when reusing items, a paired statistical analysis comparing first use and further uses of an item was conducted through a Wilcoxon's signed rank test.

The study was approved in advance by the Data Protection Officer of the TiHo. All data used and collected were analyzed and processed anonymously.

RESULTS

Establishing the Exam Processes for Electronic Examinations

The basis for the implementation of quality assurance measures, or "QA measures" for short, such as the integration of

blueprints and a multi-stage review process, took place with the system changeover to Q-Exam Institution in 2017. The examination process of electronic exams is divided into three phases: (a) conception and pre-review, (b) execution, and (c) post-review and manipulation, see Figure 2. Accompanying this, a training concept for item authors was developed that addresses pre- and post-review measures. Workshops and training materials cover topics such as working in the exam management platform, question formats and requirements, clueing issues, content review, and post-review including item and distractor analysis.

Initially, a blueprint is created for each exam, in which rules are set that define the exact number of items from each topic of the subjects. Such blueprints are created by the responsible institutes or clinics and are binding for the compilation of the examination by the exam coordinator from the items previously positively evaluated in the review process.

After the exam has been performed, the system automatically evaluates the results so that the preliminary test results and statistical parameters for the tests and items are available at the same time. In addition to the calculation of the characteristic values, a distribution analysis of the response behavior is created for the individual items, which outputs the results of the entire examination group as well as those of the top 20% and bottom 20% of the cohorts. These characteristic values are carefully checked by the exam coordinator, among other things the exam is examined with regard to reliability using Cronbach's α as well as items with conspicuous characteristic values, which are eliminated if necessary. This is followed by the initial release of the results with the possibility for the students to submit comments on individual items for a limited period of time, which are then answered by the respective authors with a statement within a certain timeframe. After expiration of the deadline, the exam coordinator carries out the post-review, which evaluates the commented items and, if necessary, manipulates the examination by removing ineffective items from the scores, whereby the results are recalculated. Afterwards, the exam results are finally released.

The results of the evaluation of exams and items are presented below.

Influence of Item Count

In Figure 3, trend analysis is used to show the average Cronbach's α values of all electronic exams from 2008 to 2022 depending on the respective number of items. Exams that include a higher number of items achieve a higher Cronbach's α. The trend analysis shows that the recommended α of .8 can be reliably reached and surpassed with 98 items per exam.

Evaluation of Key Figures

The comparative analysis of all examinations in the old Q [kjuː] examination system and all exams in the new Q-Exam Institution examination system is shown in Figure 4. Here, the reliability of all individual examinations (N=294) before and after implementation of the QA measures is presented using the Cronbach's α value. The mean Cronbach's α values are .7457 for the exams from 2008 to 2016 of the "Before Implementation" section (n=161) and .7487 for the exams from 2017 to 2022 of the "After Implementation" section (n=133). When analyzing whether the system change led to a change in exam reliability, no significant difference was found between the exams from the old system Q [kjuː] and the new system Q-Exam Institution, the p value is .5533.

Item Formats

The average difficulty of all items studied (N=12,405) was determined separately by item format. Picture mapping items (n=33) were on average the easiest with a difficulty of 82.52, followed by MCQ Type A4 (n=3,324) with 78.78, Key Feature (n=391) with 75.80, MCQ Type A3 (n=5,521) with 74.66, MCQ Type A5 (n=2,338) with 73.38, and Picture diagnosis questions (n=59) with a difficulty of 69.43, while Kprim items (n=1,039) were the most difficult with an average of 64.80. The overall value for all MCQ Type A items (n=10,883) was 75.64 and was ranked between Key Feature and Picture diagnosis.

Figure 5 shows the percentage of all items of the corresponding question formats in the respective categories of the characteristic value "difficulty."

With regard to the discrimination index, the average was also determined separately according to item formats for all items examined (N=12,405). The MCQ Type A5 format (n=2,338) separated best with a mean discrimination index of 0.287, followed by Picture mapping (n=33) with 0.246, MCQ Type A4 (n=3,324) with 0.243, and MCQ Type A3 (n=5,221) with 0.229 as well as Kprim (n=1,039) with 0.220. Picture diagnosis questions (n=59) had a discrimination index of 0.208, whereas key feature items (n=391) discriminated worst with a mean discrimination index of 0.180. The overall value for all MCQ Type A items (n=10,883) was 0.246 and corresponded to the Picture mapping format.

Regarding further statistical analysis of the differences of discriminatory properties of the item formats the Kruskal–Wallis test proves that there are significant differences in discrimination indexes of the formats (p<.0001). Subsequently assessing the differences between the individual formats illustrate that not all formats differ significantly. All statistically significant differences are shown in Table 2. To summarize the findings, it can be noted, that MCQ Type A items generally discriminate better than Kprim and Key Feature while the MCQ Type A5 format discriminates better than Picture diagnosis. Furthermore, Kprim has a significantly better discrimination index than Key Feature. There are no statistically verifiable differences between the Picture mapping format and other formats.

For a more detailed examination of the distribution of the discrimination index, Figure 6 shows the percentage of all items of the corresponding item formats in the respective categories of the characteristic value.

To better evaluate the discrimination distributions and their differences, a correlation analysis of the number of items in examinations and the discrimination index of assigned items (n=18,142) was performed. The calculation shows a slightly positive (r_s=0.0282) and significant (p=.0001) correlation between number of items in an exam and item discrimination.

With the introduction of the new examination management system, the Kprim question format was available for usage in electronic examinations. Table 2 shows the distribution of correct and incorrect answer options for the Kprim tasks considered (N=332), plus their respective percentages and average item scores.

Figure 7 depicts the distribution of characteristic values in the categories "difficulty" (Panel A) and "discrimination index" (Panel B) of the Kprim items with the respective number of answer options classified as "correct" (R).

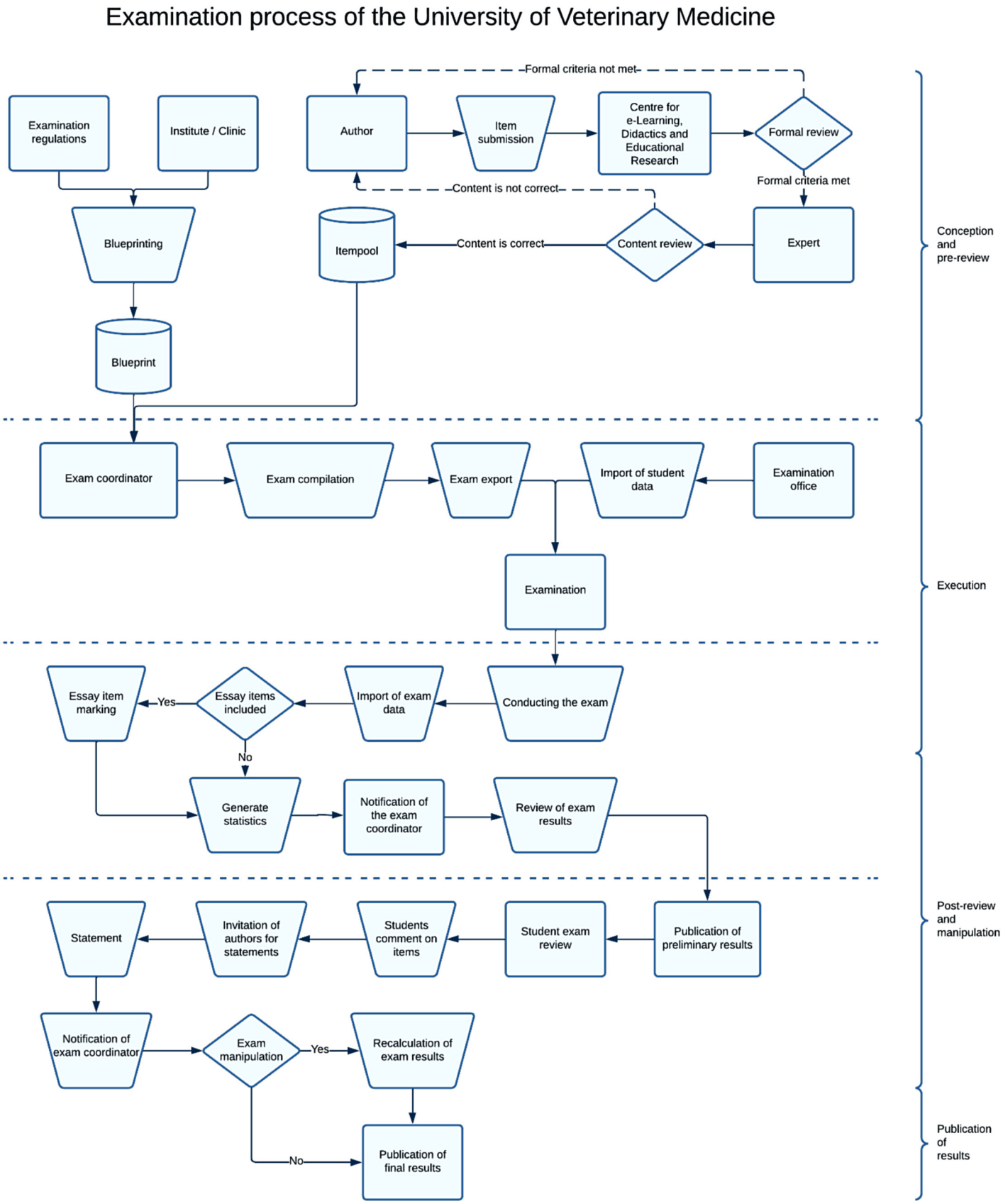

Figure 2: Examination process of electronic examinations via the examination management platform Q-Exam Institution at the University of Veterinary Medicine Hannover, Foundation.

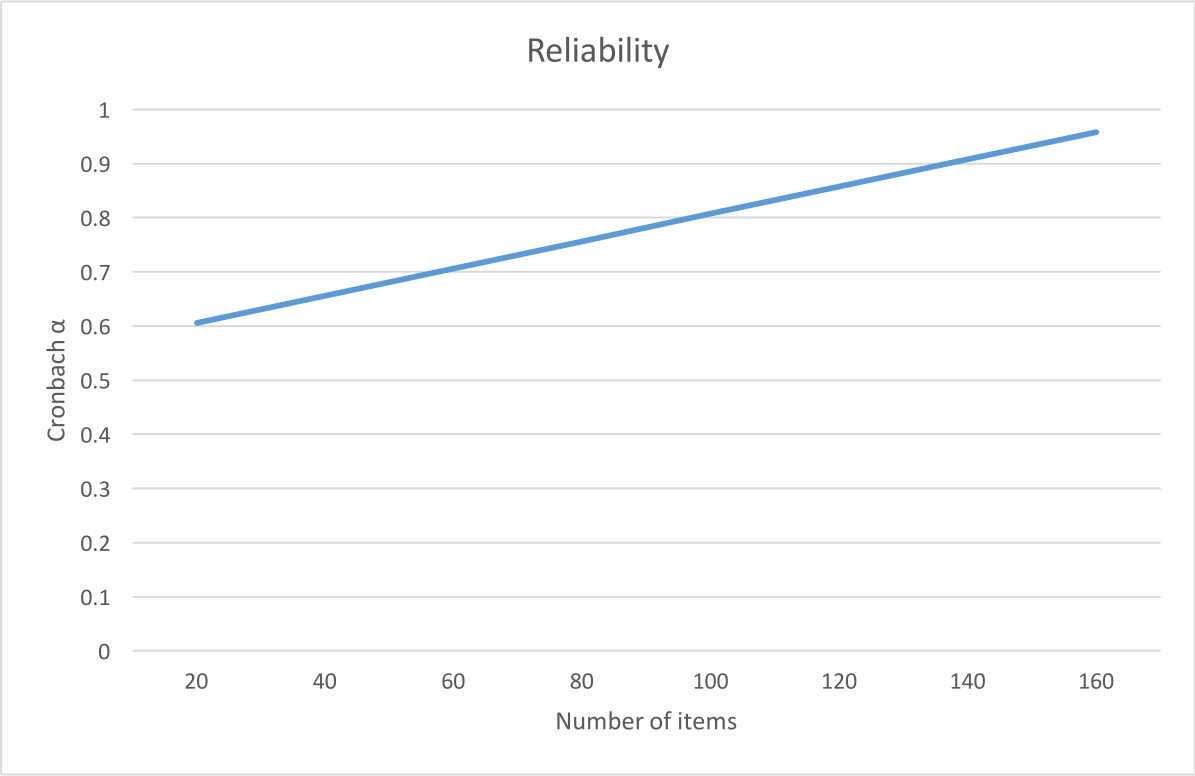

Figure 3: Trend analysis of the average Cronbach's α values of the electronic exams at the University of Veterinary Medicine Hannover, Foundation from 2008 to 2022 depending on the number of items used in the respective exams; *N*=294.

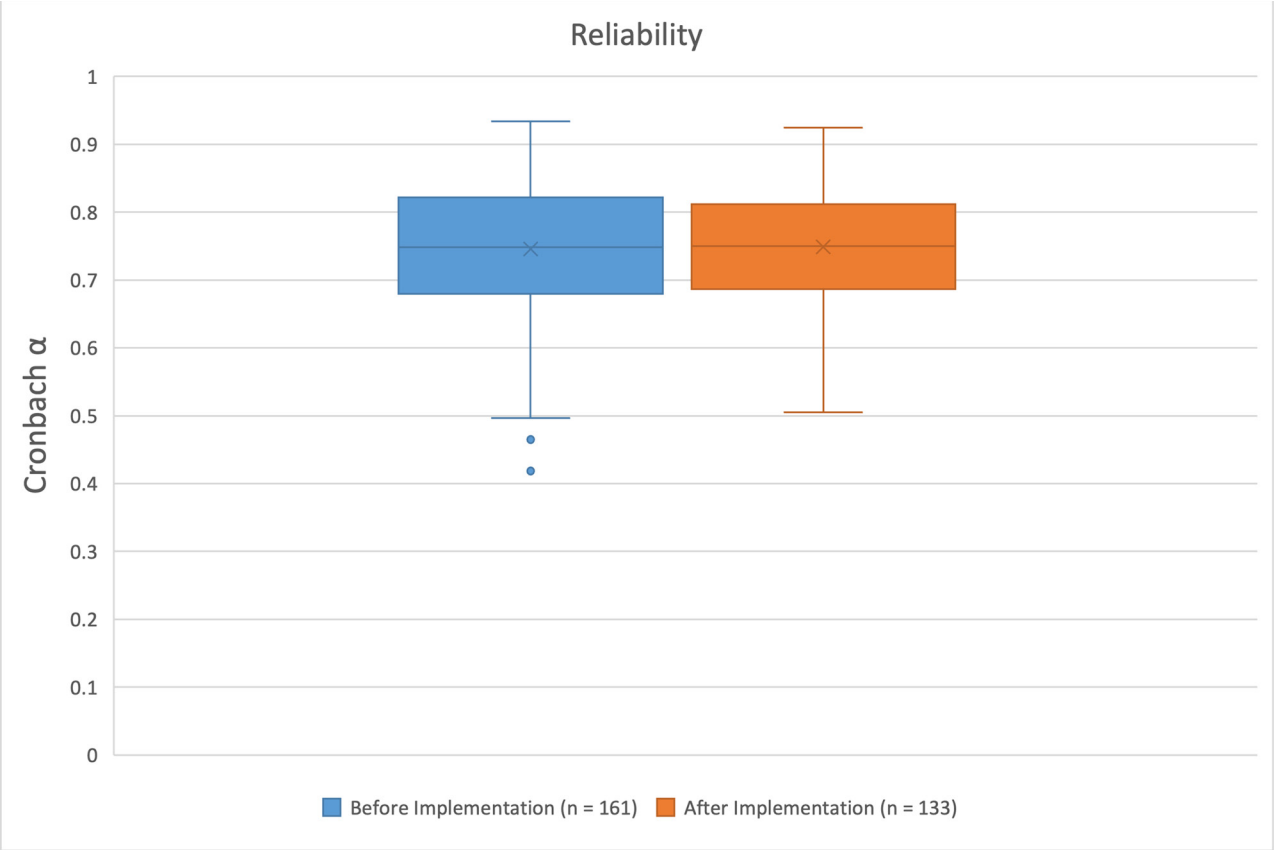

Figure 4: Distribution of the Cronbach's α values of all electronic examinations at the University of Veterinary Medicine Hannover, Foundation from 2008 to 2022, divided into the sections before implementing the QA measures in the period 2008–2016 (blue, *n*=161) and after implementing the QA measures in the period 2017–2022 (orange, *n*=133), *N*=294.

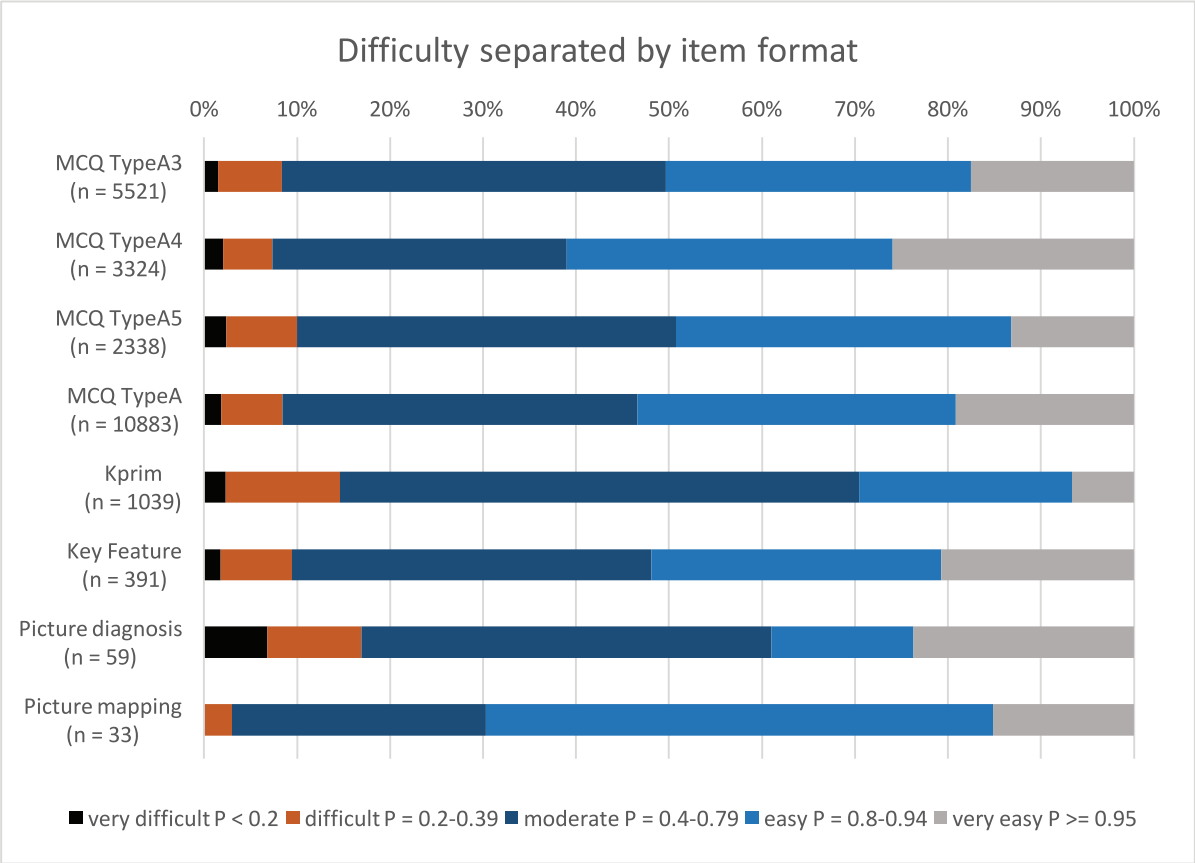

Figure 5: Relative number of items in the categories of the item characteristic "difficulty": "very difficult" (p<.2, black), "difficult" (p=.2–.39, orange), "moderate" (p=.4–.79, dark blue), "easy" (p=.8–.94, blue), and "very easy" (p≥.95, gray), separated by item format. The numbers following the term "MCQ Type A" represent the available number of choice options of these MCQ items, with "MCQ Type A" being an overview of all MCQ items regardless of the number of choice options; N=12,405.

Table 2: Significant differences in discrimination index of individual item formats using the Dwass–Steel–Critchlow–Fligner method; N=12,405.

Format comparison	Difference of discrimination indexes	p-value
MCQ Type A3 – MCQ Type A4	−0.014	<.0001
MCQ Type A3 – MCQ Type A5	−0.058	<.0001
MCQ Type A3 – key feature	0.049	<.0001
MCQ Type A4 – MCQ Type A5	−0.044	<.0001
MCQ Type A4 − Kprim	0.023	<.0001
MCQ Type A4 – key feature	0.063	<.0001
MCQ Type A5 – Kprim	0.067	<.0001
MCQ Type A5 – key feature	0.107	<.0001
MCQ Type A5 – picture diagnosis	0.079	.0016
Kprim – key feature	0.040	<.0001

Reused Items

From the pool of 7,837 items, 2,307 reused questions were identified, corresponding to a relative proportion of 29.44% of all items examined.

One focus of the evaluation was on the time interval between each use of the items in exams, looking at the relative proportion of reused questions with the respective time interval in years. The range was 12years; the average time between first use of an item and further use in a subsequent exam was about 2years (mean=2.09years). In addition to the distribution of reused questions in terms of the time interval between their following use in an exam, the effects of re-selecting an old question were evaluated separately by interval in years. On average, items became easier when used again, with an increase in difficulty of 9.69, whereas the discrimination index decreased marginally with an average change of −0.011. Paired statistical analysis of first and next usage of an item shows that the recorded changes of difficulty (p<.0001) and discrimination index (p=.0252) are statistically significant. The values for the change in item difficulty depending on the time interval in years to the next usage fluctuated in a range from –3.27 to 35.46 and showed no trend in the development. The changes in item discrimination as a function of time interval in years to next usage ranged from –0.08 to 0.09 and also showed no discernible trend.

Figure 8 presents the relationship between the proportion of reused items in individual exams and calculated overall difficulty for a sample of 22 exams for five exam subjects in the time period 2017 to mid-2022. As the proportion of reused questions increased, an exam became easier.

Moreover, the distributions of the reused items to the five question formats were examined with regard to the total

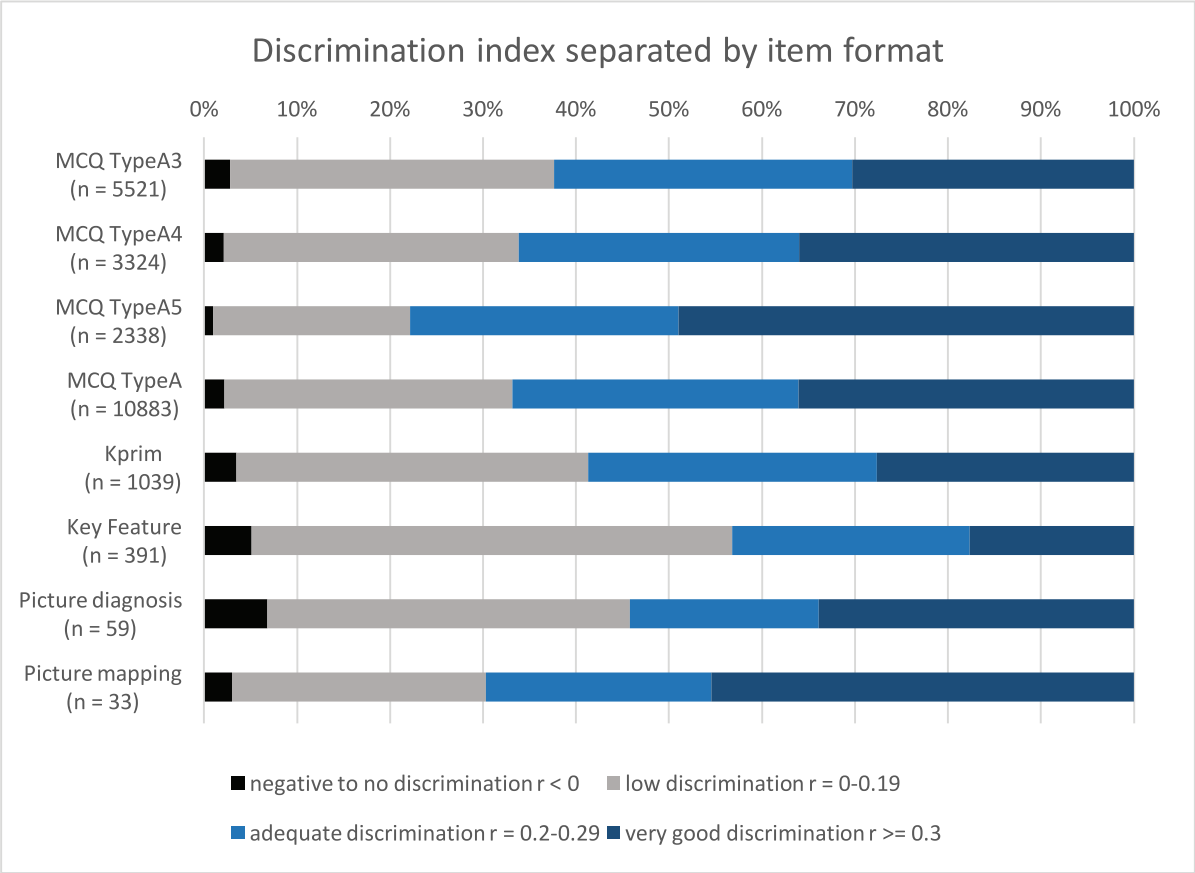

Figure 6: Relative number of items in the categories of item discrimination "negative to no discrimination" ($r<0$, black), "low discrimination" ($r=0–0.19$, gray), "adequate discrimination" ($r=0.2–0.29$, blue), and "very good discrimination" ($r\geq0.3$, dark blue), separated by item format. The numbers following the term "MCQ Type A" represent the available number of choice options of these MCQ items, with "MCQ Type A" being an overview of all MCQ items regardless of the number of choice options; $N=12,405$.

Table 3: Relative proportion of Kprim items with corresponding number of correct answer options among all Kprim items as well as their average characteristic values of "difficulty" and "discrimination index", $N=332$.

Number of right answers	Proportion of all Kprim items	Mean difficulty	Mean discrimination index
0	0.60%	81.69	0.32
1	4.52%	46.75	0.22
2	51.51%	61.74	0.20
3	34.64%	59.70	0.16
4	8.73%	54.47	0.17

number of the reused questions and to all examined items of the respective item types, whereupon no assignment of a format was possible for 103 items, of which 84 were identified as reused questions. Thus, the number of examined items was reduced to 7,734 and the number of reused questions to 2,223 items.

The largest proportion of reused items examined ($n=2,223$) consisted of MCQ Type A tasks at 82.95%, followed by Kprim at 13.68%, Key Feature at 2.70%, Picture diagnosis at 0.45%, and Picture mapping at 0.22%. In relation to all examined items ($n=7,734$) of the respective item formats, 28.78% of all MCQ Type A items ($n=6,408$) were reused as well as 29.34% of the Kprim questions ($n=1,036$), 27.91% of the Key Feature format ($n=215$), 22.73% of the Picture diagnosis items ($n=44$), and 16.13% of the Picture mapping tasks ($n=31$).

DISCUSSION

Establishing the Examination Processes for Electronic Examinations

The study was intended to be the first comprehensive review of the quality characteristics as well as the achievement of the standards and implementation of the recommendations of Jünger and Just[6] with particular reference to the changeover to the new examination management system. With the system changeover, quality assurance processes were implemented, including the integration of blueprints, training for question authors, and a formal as well as content-related pre- and post-review, which were already identified in the literature as the most important measures for ensuring and improving examination quality.[6,15,34] Blueprints ensure that the examination content is consistent with the curriculum,[35] which is considered one of the main aspects of validity.[8] The focus, however, is on the optimization of the items, since these essentially determine the quality of the examinations in general.[2] Studies could already prove that suitable training concepts for item authors as well as the structured, multi-stage

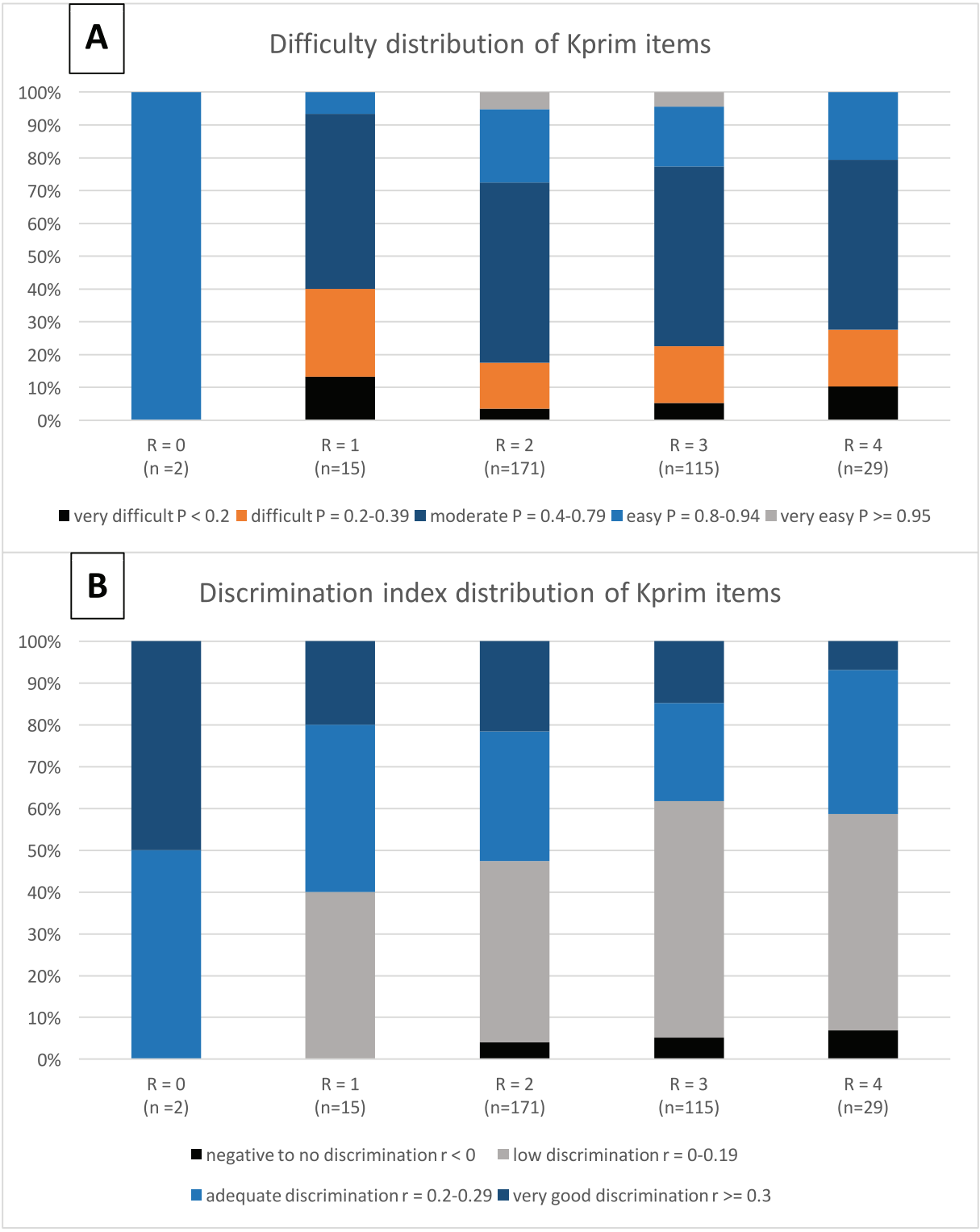

Figure 7: Panel A: Relative number of Kprim items in the categories of the item characteristic "difficulty": "very difficult" (p<.2, black), "difficult" (p=.2–.39, orange), "moderate" (p=.4–.79, dark blue), "easy" (p=.8–.94, blue), and "very easy" (p≥.95, gray). Panel B: Relative number of Kprim items in the categories of the item characteristic "discrimination index": "negative to no discrimination" (r<0, black), "low discrimination" (r=0–0.19, gray), "adequate discrimination" (r=.2–.29, blue), and "very good discrimination" (r≥0.3, dark blue); N=332.

review process not only increase the efficiency of item design[36] but also have a significant positive impact on item quality and reliability,[2,15,36–38] which underlines the relevance of these measures. At the TiHo, the aforementioned quality assurance measures were successfully established and firmly integrated into the examination process, which at the same time created a continuous exchange between all those involved in assessment procedures. Additionally, the range of training courses and materials for self-study is continuously being improved and expanded.

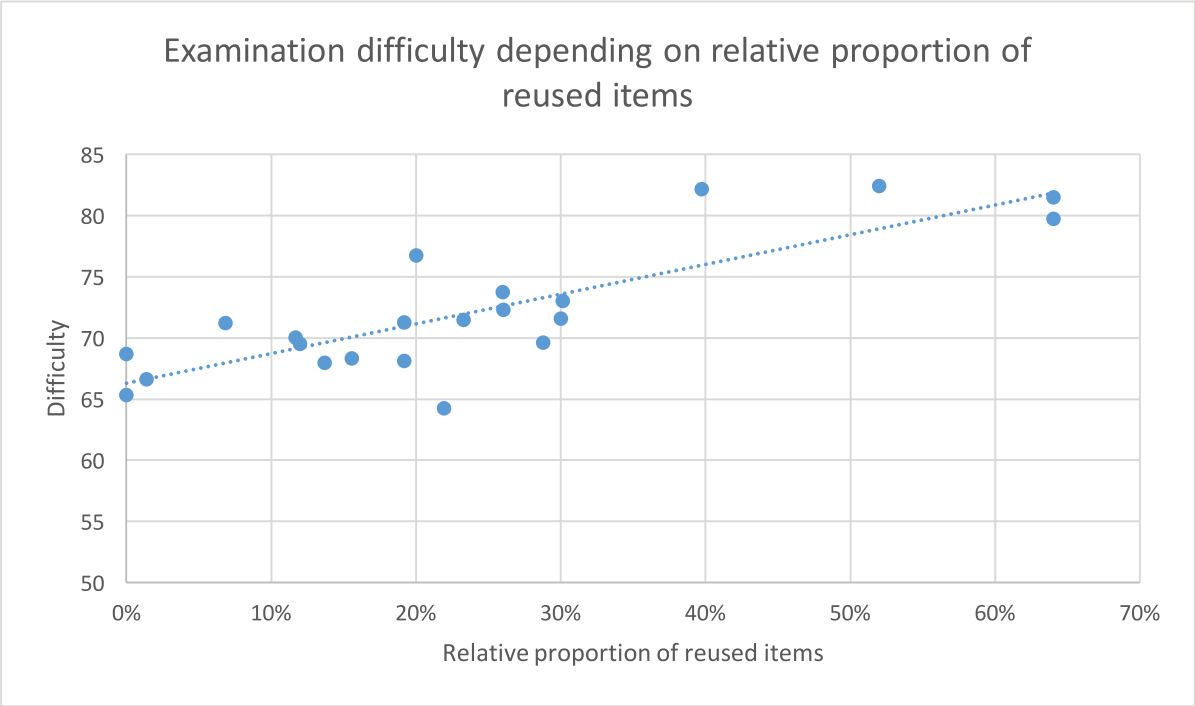

Figure 8: Examination difficulty of the individual exams depending on the relative proportion of reused items in the respective exam with trend line; *N*=22.

Influence of Item Count

Reliability is generally regarded as an important quality criterion for the assessment of written exams, and even tends to be the most important one, since, in contrast to the other two quality criteria of objectivity and validity, it can be measured best and thus be used as a quantifiable characteristic for evaluating and comparing tests.[19] In the literature, item quality and item number,[2,6,13,15,17,19] are commonly considered to be factors influencing reliability. The quality of the items is ensured by the training concepts and review processes already discussed, but cannot be statistically recorded and assessed before the items are used for the first time, making it difficult to use item quality in practice to improve reliability before the examination is carried out. As an alternative, there is the possibility to eliminate items in the post-review process afterwards. The effect of individual items on the reliability can be calculated by analyzing the Cronbach's α while excluding the chosen item. Thus, items with a negative effect on reliability can be identified and removed from scoring in order to improve overall reliability of the exam in retrospect. However, the focus for achieving the necessary threshold value for Cronbach's α of .8 before the examination is performed is on the use of an adequate number of items in the exams. In this regard, Jünger and Just[6] state a minimum number of 40 high-quality questions, whereas Kibble[15] recommends 50–60 items for items with an average good discrimination index ($r \geq 0.3$) and 100 items for an average adequate discrimination index ($r=0.2$), while Krebs[2] considers up to 300 questions necessary for extensive subjects, such as clinical medicine. The results of this study support these statements, as the minimum reliability of 0.8 in the TiHo examinations was only safely reached from a number of 98 items and shows the positive correlation of number of items and the Cronbach's α value. It is therefore recommended to constantly monitor the test reliability of all exams by calculating and evaluating the Cronbach's α and, if necessary, to adjust the number of items in case of inadequate values.

Evaluation of Key Figures

The evaluation of the examination reliability on the basis of Cronbach's α makes it clear that the implementation of the quality assurance measures with the system change at the TiHo has only led to a non-significant improvement in the reliability.

A possible influencing factor for this result is the fact that a large proportion of the authors of questions had already developed items for examinations prior to the system change and had participated in internal training courses for authors beforehand, thus gaining experience in creating good items, so that the new training concepts and the pre-review did not have such a significant effect on examination and item quality as could be determined in other studies.[15,36–38]

Item Formats

Taking into consideration the mean characteristic values as well as characteristic value distributions, the MCQ Type A5 format was qualitatively better than multiple choice items with three or four choice options, which is contrary to the general consensus in the literature that three choice options are the optimal number.[2,7,39–42] Regarding this, however, it must be borne in mind that all MCQ Type A5 items come from examinations with a very high number of items, so they are examinations with a correspondingly high Cronbach's α due to the positive correlation between item number and reliability discussed earlier, which means that the discrimination index is generally higher for such items both for the correlation between items[36] and for the correlation between reliability and discrimination index.[43] This characteristic was proven through statistical analysis in that the discrimination index of items from exams with a high number of questions tends to be better. Due to the fact that MCQ Type A3 and MCQ Type A4 are

also increasingly being used in exams with a low item count, this result must be considered critical.

Regarding the Kprim format, it has been found in the literature that Kprim items with half-point scoring, as practiced at the TiHo, have better values regarding the psychometric parameters difficulty and discrimination index than MCQ Type A tasks.[44] These observations cannot be confirmed in the present study, since Kprim represents the most difficult of all investigated formats and shows a worse discrimination index in comparison than MCQ Type A formats and Picture mapping. Especially regarding the correlation between difficulty and discrimination index, this is considered an unexpected result, since items with an intermediate difficulty index are supposed to have the best discrimination index values,[2,45–47] with optimal ranges between 40%–60%[29,46] and 40%–74%[45] being reported, and deviations to a lower or higher difficulty index are supposed to result in a significantly lower discrimination index. Since the Kprim tasks with an average difficulty of 64.80% were significantly closer to a moderate difficulty than, for example, the MCQ Type A items with 73.38%–78.78% and, with respect to the distribution analysis, had the largest proportion of items within a moderate difficulty (p=40%–79%) with 55.82%, it would be expected in theory that the Kprim format would show a better discrimination index than the other formats, but this is not the case in the evaluation. A possible influencing factor is the lack of heterogeneity in the number of correct answers of the items, since over 86% of all Kprim tasks were designed with either two or three correct choice options, so presumably students who might be aware of this fact may have an increased guessing probability, which in turn has a negative impact on the psychometric parameters. Another factor is that the formulation of good Kprim tasks and choices is sometimes very difficult, which in turn affects the characteristic values.[2,7] Regardless of this, the exact background of this result cannot be evaluated on the basis of the study. Furthermore, the Picture mapping format stands out as an adequate item format due to the discrimination index analysis, but the sample for this task format was comparatively small with 33 questions.

In contrast, the format Key Feature has to be considered more closely. Although the difficulty distribution is comparable to MCQ Type A3 and MCQ Type A4, Key Feature had the worst discrimination index distribution among all item formats with a mean discrimination index of 0.18, this being in an unacceptable range. Currently, there is no literature on the metrological evaluation of Key Feature items as they are used at the TiHo, which means that no statement can be made at this stage as to whether this is a general issue with the question format or whether internal university factors have the greatest influence on the quality of this format.

In summary, the MCQ Type A format at the TiHo turns out to be the best format in terms of measurement in accordance with the literature[15,48] and should therefore continue to be used primarily in examinations. The Picture mapping format also stands out with a good discrimination index and should therefore be used more frequently in the future, but should be closely monitored due to the small size of the sample. Contrary to expectations, Kprim and key feature items stand out as poorer item formats in relation to the others, which is why these formats should be examined more closely in the review process in the future.

Reused Items

The constant conception of high-quality items for the annual examinations is a time-consuming and labor-intensive task for the authors of questions.[49] In this context, the creation of a question database using an electronic exam management platform can be used to build an item pool of tasks with adequate psychometric parameters that can be reused in later exam cycles. However, the reuse of questions bears the risk that students will collect and pass on items to subsequent exam cohorts[50–52] that can be incorporated into the learning process. It stands to reason that knowledge of items can have an impact on quality, which is why this study was used to examine the effects on question characteristics of the difficulty and discrimination indexes. Thereby, it could be determined that the reuse of items partly showed a strong influence on the difficulty and the questions became easier with an increase of 9.69% on average, but the discrimination index in turn decreased only minimally. This result is reflected in other studies, which also found a clear change in the difficulty index and only marginal deviations in the discrimination index.[49–51,53] In one previous study even a positive effect on the discrimination index was observed.[54]

Moreover, trends or correlations could be derived within the framework of these aforementioned studies, since the changes in the characteristic values were only significant from the second reuse of the items[50,53] and could only be detected marginally from a time interval of 5years.[50,51] Such statements cannot be supported with this present study, since these developments were not detectable.

Most importantly, the observed effect on item difficulty must be considered when compiling exam questions because, as can be seen in Figure 8, the percentage of students who were able to choose the correct answers increases as the proportion of reused items increases, so the exams become easier overall. As a consequence, examinees' performance might be falsely overestimated, which should be avoided especially in career-determining state exams. One possible solution to avoid overestimation of examinee performance is to apply a criterion-referenced standard setting for written examinations, such as the three-level Angoff method,[55] in order to set an appropriate passing cut score for each specific examination. For universities that have already introduced a fixed cut score and anchored it in the examination regulations, this could potentially prove difficult to implement.

Reusing items is reasonable and can provide many benefits, one of them being the reuse of items with good characteristic values, which is why the reused questions should have as good characteristic values as possible.[49] On the other hand, the results of this study and those of other evaluations show that the proportion of reused items in examinations should be kept low[56] at around 20% of all items—because these items might be recognized by examinees. It is recommended that reused questions should only be used once again in an examination[50,53] and that the time interval to the last use should be at least 3–5years.[50,51]

With regard to the proportions of reused questions in the item formats of all items examined, it turns out that with the relatively most frequently occurring formats MCQ Type A, Kprim and Key Feature there is no tendency to increased reuse of one of these formats.

Limitations and Implications of the Study

The most common test theory models to interpret examination data are classical test theory (CTT) and item response theory (IRT). CTT, which was used in this study, has some drawbacks, the most important one being dependent on the

examinee sample and item sample.[57–59] However, there are also several aspects rendering the more mathematically complex IRT models impracticable or even not realistically feasible for smaller institutions. Firstly, IRT models are reliant on accurate estimates of item parameters and model fit, which can be costly and difficult to realize.[60] In order to calculate these estimates a minimum of at least 500 examinees is needed, whereas a minimum of 200 candidates is sufficient for CTT models.[58] Taking the TiHo as an example with an average of 248 examinees, the number of participants of each exam are simply not adequate to calibrate item parameter estimates with IRT models, especially for smaller veterinary medicine faculties. Furthermore, since IRT models are more complex and demanding[58] trained personnel is needed for application of IRT models, which might not be available for smaller institutions. In contrast, CTT is easily understandable, and the user does not require much mathematical knowledge.[58,59] By carefully studying item statistics, CTT is easy to use[58] for exam coordinators after an introductory training. For most medical education settings with locally used assessments, CTT is considered to be an adequate method.[59] In conclusion it is recommended for smaller faculties to implement CTT as the preferred method of analyzing written examinations.

While this study was conducted based on an examination process that is specific to the TiHo, this process, which has already been established for years, and the measures for evaluating the exams can serve as an example on the basis of which ideas for implementing or improving one's examinations process can be derived. One measure that is generally applicable and easy to implement are offers to support those involved in the process. This includes consultations, tutorials, regular training sessions and workshops for question authors, content reviewers and exam coordinators, for whom different topics are of interest.

At the TiHo, question authors and content reviewers are trained primarily in the use of the exam management platform and on formal and content requirements for exam questions. Such criteria for items include relevance of the items to the intended learning objectives, alignment of difficulty with the competence of a newly starting professional, grammar and comprehensibility of items, answerability without seeing the answers, homogeneity of answer options in terms of topic area, length and detail, plausibility of distractors, and assessment of taxonomy level. In addition, authors and reviewers are trained to recognize solution clues in items (e.g., grammatical cues, logical cues, absolute terms or convergence strategy[61]). These requirements are based on the recommendations for creating exam items for written assessments from international literature, such as the publication by Case and Swanson.[61]

Furthermore, exam coordinators are advised on the creation of blueprints as well as on the selection of appropriate item formats for the exam and their ratio. Aspects such as distribution of taxonomy levels of items, assessment of item difficulty, reuse of items, adequate ratio of easy and difficult items in exams, and selection of an appropriate number of items for the subject are also discussed. In addition, it is considered important to provide exam coordinators with tools for an effective post-review of exams and items. For this purpose, workshops on exam, item and distractor analysis are offered, including the quality criteria that were evaluated in this study.

Generally speaking, the examination process, the exams and their quality should be constantly monitored in order to identify opportunities for improvement, increased efficiency and further development. Support and exchange offers should be implemented for all staff responsible for the exams in order to increase the competence of those involved and to promote an open discourse.

CONCLUSIONS

The results of this study show and underline the relevance of measures accompanying examinations, such as pre- and post-review processes, in order to implement electronic examinations at a consistently high level. Based on the results, the first step was to increase the number of items in examinations that consistently failed to achieve a Cronbach's α value >.8, in consultation with those responsible for the examinations, and to adapt the examination regulations accordingly. Attention was also paid to the reused items and their proportion in the examination composition.

Furthermore, the distribution of the examination results, reliability, item analysis statistics including the parameters difficulty index and discrimination index as well as distractor analysis were identified as suitable indicators for the evaluation of the examinations and items for everyday practice, communicated with the persons responsible for the examinations, and stored in a checklist for the post-review process.

Based on the results of the study, recommendations were thus developed for examination coordinators and lecturers for their own internal quality assurance, and indicators were identified that will be considered and evaluated on an ongoing basis in order to be able to continuously ensure and improve the quality of examinations.

ETHICAL RECOGNITION

This study was conducted according to the ethical standards of the University of Veterinary Medicine Hannover, Foundation. The doctoral thesis committee of the University, which acts as the University's ethics committee, validated the project in accordance to ethical guidelines regarding research with human participants and approved the study.

AUTHOR CONTRIBUTIONS

RR and ES conceived and designed the study. RR collected the data, analyzed and interpreted them, supervised by ES. The manuscript was written by RR with critical input and notable revisions from ES and AT. All authors reviewed and approved the final version.

CONFLICTS OF INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ACKNOWLEDGMENTS

This publication was supported by Stiftung Innovation in der Hochschullehre within the project "FERVET – Digital Teaching and Review of Clinical Practical Skills in Veterinary Medicine from an Animal Welfare Perspective."

REFERENCES

1 Schuwirth LWT, van der Vleuten CPM. How to design a useful test: the principles of assessment. In: Swanwick T, Forrest K, O'Brien BC, editors. Understanding medical education: evidence, theory and practice. 3rd ed. Hoboken, NJ: Wiley-Blackwell; 2018. p. 275–89.

2 Krebs R. Prüfen mit multiple choice. Kompetent planen, entwickeln, durchführen und auswerten. 1st ed. Bern: Hogrefe; 2019.
3 Norcini JJ, Shea JA. The credibility and comparability of standards. Appl Meas Educ. 1997;10(1):39–59. https://doi.org/10.1207/s15324818ame1001_3
4 Kane MT, Crooks TJ, Cohen AS. Designing and evaluating standard-setting procedures for licensure and certification tests. Adv Health Sci Educ Theory Pract. 1999;4(3):195–207. https://doi.org/10.1023/A:1009849528247. PMID: 12386478
5 Schüttpelz-Brauns K, Schubert S. Qualitätssicherung von Multiple-Choice-Prüfungen. In: Dany S, Szczyrba B, Wildt J, editors. Prüfungen auf die Agenda. Bielefeld: W. Bertelsmann; 2008. p. 92–102.
6 Jünger J, Just I. Recommendations of the German Society for Medical Education and the German Association of Medical Faculties regarding university-specific assessments during the study of human, dental and veterinary medicine. GMS Z Med Ausbild. 2014;31(3):1–23. https://doi.org/10.3205/zma000926. PMID: 25228936
7 Jolly B, Dalton MJ. Written assessment. In: Swanwick T, Forrest K, O'Brien BC, editors. Understanding medical education: evidence, theory and practice. 3rd ed. Hoboken, NJ: Wiley-Blackwell; 2018. p. 291–317.
8. Tavakol M, Dennick R. The foundations of measurement and assessment in medical education. Med Teach. 2017;39(10):1010–15. https://doi.org/10.1080/0142159X.2017.1359521. PMID: 28768456
9 Dennick R, Wilkinson S, Purcell N. Online eAssessment: AMEE guide no. 39. Med Teach. 2009;31(3):192–206. https://doi.org/10.1080/01421590902792406. PMID: 19811115
10 Van Der Vleuten CPM. The assessment of professional competence: developments, research and practical implications. Adv Health Sci Educ Theory Pract. 1996;1(1):41–67. https://doi.org/10.1007/BF00596229. PMID: 24178994
11 Norcini J, Anderson B, Bollela V, et al. Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 conference. Med Teach. 2011;33(3):206–214. https://doi.org/10.3109/0142159X.2011.551559. PMID: 21345060
12 Downing SM. Assessment of knowledge with written test forms. In: Norman GR, van der Vleuten CPM, Newble DI, Dolmans DHJM, Mann KV, Rothman A, et al., editors. International handbook of research in medical education. Dordrecht: Springer Netherlands; 2002. p. 647–72.
13 Downing SM. Reliability: on the reproducibility of assessment data. Med Educ. 2004;38(9):1006–12. https://doi.org/10.1111/j.1365-2929.2004.01932.x. PMID: 15327684
14 Dent JA, Harden RM. A practical guide for medical teachers. London: Churchill Livingstone; 2001.
15 Kibble JD. Best practices in summative assessment. Adv Physiol Educ. 2017;41(1):110–119. https://doi.org/10.1152/advan.00116.2016. PMID: 28188198
16 Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika. 1951;16(3):297–334.
17 Tavakol M, Dennick R. Making sense of Cronbach's alpha. Int J Med Educ. 2011;2:53–55. https://doi.org/10.5116/ijme.4dfb.8dfd. PMID: 28029643
18 McKinley RK, Fraser RC, Baker R. Model for directly assessing and improving clinical competence and performance in revalidation of clinicians. BMJ. 2001;322(7288):712–715. https://doi.org/10.1136/bmj.322.7288.712. PMID: 11264212
19 Möltner A, Schellberg D, Jünger J. Grundlegende quantitative Analysen medizinischer Prüfungen. GMS Z Med Ausbild. 2006;23(3):11.
20 DeVellis RF. Scale development: theory and applications. Thousand Oaks, CA, USA: Sage Publications, Inc., 1991. p. 121.
21 McCowan RJ, McCowan SC. Item analysis for criterion-referenced tests. New York: Center for Development of Human Services; 1999.
22 Kuhn S, Frankenhauser S, Tolks D. Digitale Lehr- und Lernangebote in der medizinischen Ausbildung. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz. 2018;61(2):201–209. https://doi.org/10.1007/s00103-017-2673-z
23 Brahm T, Seufert S, editors. "Ne(x)t generation learning". E-assessment und E-portfolio: halten sie, was sie versprechen? St. Gallen: SCIL, Swiss Centre for Innovations in Learning; 2007.
24 Gruttmann SJ. Formatives E-assessment in der Hochschullehre – Computerunterstützte Lernfortschrittskontrollen im Informatikstudium. Paderborn: MV-Verlag; 2010.
25 Ehlers JP, Guetl C, Höntzsch S, Usener CA, Gruttmann SJ. Prüfen mit Computer und Internet. Didaktik, Methodik und Organisation von E-Assessment. In: Ebner M, Schoen S, editors. L3T Lehrbuch für Lernen und Lehren mit Technologien. 2nd ed. Berlin: epubli, 2013.
26 Ellaway R, Masters K. AMEE guide 32: e-learning in medical education. Part 1. Learning, teaching and assessment. Med Teach. 2008;30(5):455–73. https://doi.org/10.1080/01421590802108331. PMID: 18576185
27 Clauser BE, Schuwirth LWT. The use of computers in assessment. In: Norman GR, van der Vleuten CPM, Newble DI, Dolmans DHJM, Mann KV, Rothman A, et al., editors. International handbook of research in medical education. Dordrecht: Springer Netherlands; 2002. p. 757–92.
28 Thorndike RM, Cunningham GK, Thorndike RL, Hagen EP. Measurement and evaluation in psychology and education, 5th ed. New York, NY, England: Macmillan Publishing Co., Inc.; 1991. p. 544.
29 Hermi A, Achour W. Item analysis of examinations in the Faculty of Medicine of Tunis. Tunis Med. 2016;94(4):247–52. PMID: 27704506
30. Ebel RL, Frisbie DA. Essentials of educational measurement. 5th ed. Englewood Cliffs, N.J.: Prentice-Hall; 1986.
31 Ehlers JP, Carl T, Windt K-H, Möbs D, Rehage J, Tipold A. Blended assessment: Mündliche und elektronische Prüfungen im klinischen Kontext. Zeitschrift für Hochschulentwicklung. 2010;4(3):24–36.
32 Schaper E, Ehlers JP. 6 Jahre eAssessment an der Stiftung Tierärztliche Hochschule Hannover. Hamburger Elearning Magazin. 2011;(7):43–44.
33 Raymond MR, Grande JP. A practical guide to test blueprinting. Med Teach. 2019;41(8):854–61. https://doi.org/0.1080/0142159X.2019.1595556. PMID: 31017518
34 Schurter T, Escher M, Gachoud D, et al. Essential steps in the development, implementation, evaluation and quality assurance of the written part of the Swiss federal licensing examination for human medicine. GMS J Med Educ. 2022;39(4):Doc43. https://doi.org/10.3205/zma001564. PMID: 36310888
35 Notar CE, Zuelke DC, Wilson JD, Yunker BD. The table of specifications: insuring accountability in teacher made tests. J Instr Psychol. 2004;31(2):115.
36 Rotthoff T, Soboll S. Quality assurance of multiple choice questions. An exemplary way for a medical faculty. GMS Z Med Ausbild. 2006;23(3):Doc45.
37 Naeem N, van der Vleuten CPM, Alfaris EA. Faculty development on item writing substantially improves item quality. Adv Health Sci Educ Theory Pract. 2012;17(3):369–76. https://doi.org/10.1007/s10459-011-9315-2

38 Malau-Aduli BS, Zimitat C. Peer review improves the quality of MCQ examinations. Assess Eval High Educ. 2012;37(8):919–31. https://doi.org/10.1080/02602938.2011.586991

39 Hogben D. The reliability, discrimination and difficulty of word-knowledge tests employing multiple choice items containing three, four, or five alternatives. Aust J Educ. 1973;17(1):63–8. https://doi.org/10.1177/000494417301700107

40 Tversky A. On the optimal number of alternatives at a choice point. J Math Psychol. 1964;1(2):386–91. https://doi.org/10.1016/0022-2496(64)90010-0

41 Rodriguez MC. Three options are optimal for multiple-choice items: a meta-analysis of 80years of research. Educ Meas. 2005;24(2):3–13. https://doi.org/10.1111/j.1745-3992.2005.00006.x

42 Tarrant M, Ware J. A comparison of the psychometric properties of three- and four-option multiple-choice questions in nursing assessments. Nurse Educ Today. 2010;30(6):539–43. https://doi.org/10.1016/j.nedt.2009.11.002. PMID: 20053488

43 Setiawana A, Weku WCD. Simulation study of reliability coefficient and discrimination index. IConSSE FSM SWCU, 2015; Salatiga, Indonesia.

44 Lahner F-M, Nouns ZM, Krebs R, Fischer MR, Huwendiek S. Schriftliche Prüfungen: Vorteile von Multiple True-False Fragen gegenüber Typ-A Fragen. In: Gemeinsame Jahrestagung der Gesellschaft für Medizinische Ausbildung (GMA) und des Arbeitskreises zur Weiterentwicklung der Lehre in der Zahnmedizin. Düsseldorf: German Medical Science GMS Publishing House; 2015. p. 19.

45 Sim S-M, Rasiah RI. Relationship between item difficulty and discrimination indices in true/false-type multiple choice questions of a para-clinical multidisciplinary paper. Ann Acad Med Singap. 2006;35(2):67–71. https://doi.org/10.47102/annals-acadmedsg.V35N2p67. PMID: 16565756

46 Pande SS, Pande SR, Parate VR, Nikam AP, Agrekar SH. Correlation between difficulty and discrimination indices of MCQs in formative exam in physiology. SEAJME. 2013;7(1):45–50. https://doi.org/10.4038/seajme.v7i1.149

47 Mitra N, Nagaraja H, Ponnudurai G, Judson J. The levels of difficulty and discrimination indices in type A multiple choice questions of pre-clinical semester 1 multidisciplinary summative tests. IeJSME. 2009;3(1):2–7. https://doi.org/10.56026/imu.3.1.2.

48 Krebs R. Anleitung zur Herstellung von MC-Fragen und MC-Prüfungen für die ärztliche Ausbildung. Bern: Institut für Medizinische Lehre IML, Abteilung für Ausbildungs- und Examensforschung AAE; 2004.

49 Malik RH, Malik AS. Developing a bank of faculty-authored, valid and reliable objective questions for institutional use: sharing the experience. Mal J Med Health Sci. 2020;16(7):28–35.

50 Joncas SX, St-Onge C, Bourque S, Farand P. Re-using questions in classroom-based assessment: an exploratory study at the undergraduate medical education level. Perspect Med Educ. 2018;7(6):373–8. https://doi.org/10.1007/s40037-018-0482-1. PMID: 30421331

51 Möltner A, Egarter S, Albrecht T. Verwendung von Altfragen in Prüfungen: Einfluss von Zahl der Wiederholungen und dem zeitlichen Abstand zur Letztverwendung am Beispiel der Hals-Nasen-Ohrenheilkunde der Medizinischen Fakultät Heidelberg. Gemeinsame Jahrestagung der Gesellschaft für Medizinische Ausbildung (GMA) und des Arbeitskreises zur Weiterentwicklung der Lehre in der Zahnmedizin (AKWLZ); 15.-17.09.2022; Halle (Saale). Düsseldorf: German Medical Science GMS Publishing House; 2022.

52 Lowe D, Foulkes J, Russell R. ABS to MRCS at the RCS: philosophy, format and future. Ann R Coll Surg Engl. 1998;80(5 Suppl):213–8. PMID: 10343551

53 Panczyk M, Zarzeka A, Malczyk M, Gotlib J. Does repetition of the same test questions in consecutive years affect their psychometric indicators? – five-year analysis of in-house exams at Medical University of Warsaw. EURASIA J Math Sci Tech Ed. 2018;14(7):3301–3309. https://doi.org/10.29333/ejmste/91681

54 Appelhaus S, Werner S, Grosse P, Kämmer JE. Feedback, fairness, and validity: effects of disclosing and reusing multiple-choice questions in medical schools. Med Educ Online. 2023;28(1):2143298. https://doi.org/10.1080/10872981.2022.2143298. PMID: 36350605

55 Yudkowski R, Downing SM, Popescu M. Setting standards for performance tests: a pilot study of a three-level Angoff method. Acad Med. 2008;83(10):13–16. https://doi.org/10.1097/ACM.0b013e318183c683. PMID: 18820491

56 Seifert T, Becker T, Buttcher AF, Herwig N, Raupach T. Restructuring the clinical curriculum at University Medical Center Gottingen: effects of distance teaching on students' satisfaction and learning outcome. GMS J Med Educ. 2021;38(1):Doc1. https://doi.org/10.3205/zma001397. PMID: 33659606

57 Hambleton RK. Principles and selected applications of item response theory. In: Linn RL, editor. Educational measurement. 3rd ed. London: Macmillan Publishing Co., Inc.; 1989.

58 Hambleton RK, Jones RW. Comparison of classical test theory and item response theory and their applications to test development. Educ Meas Issues Pract. 1993;12(3):38–47. https://doi.org/10.1111/j.1745-3992.1993.tb00543.x

59 De Champlain AF. A primer on classical test theory and item response theory for assessments in medical education. Med Educ. 2010;44(1):109–17. https://doi.org/10.1111/j.1365-2923.2009.03425.x. PMID: 20078762

60 Jabrayilov R, Emons WH, Sijtsma K. Comparison of classical test theory and item response theory in individual change assessment. Appl Psyc Meas. 2016;40(8):559–72. https://doi.org/10.1177/0146621616664046. PMID: 29881070

61 Case SM, Swanson DB. Constructing written test questions for the basic and clinical sciences. 2nd ed. Philadelphia: National Board of Medical Examiners; 1998.

AUTHOR INFORMATION

Robin Richter, DVM (0000-0001-7596-2747), is a doctoral student and research assistant in education research at the Centre for E-Learning, Didactics and Educational Research, University of Veterinary Medicine Hannover, Bünteweg 2, 30559 Hannover, Germany. Email: robin.richter@tiho-hannover.de.

Andrea Tipold, Dr. med. vet, DECVN (0000-0002-9421-942X), is vice president for teaching and professor of neurology, Neurology, Department of Small Animal Medicine and Surgery, University of Veterinary Medicine Hannover, Bünteweg 9, 30559 Hannover, Germany.

Elisabeth Schaper, Dr. med. vet. (0000-0002-9559-1995), is a Research Associate in digital higher veterinary education, Centre for E-Learning, Didactics and Educational Research, University of Veterinary Medicine Hannover, Bünteweg 2, 30559 Hannover, Germany.