An algorithm for calculating exam quality as a basis for performance-based allocation of funds at medical schools

Abstract

Objective: The amendment of the Medical Licensing Act (ÄAppO) in Germany in 2002 led to the introduction of graded assessments in the clinical part of medical studies. This, in turn, lent new weight to the importance of written tests, even though the minimum requirements for exam quality are sometimes difficult to reach. Introducing exam quality as a criterion for the award of performance-based allocation of funds is expected to steer the attention of faculty members towards more quality and perpetuate higher standards. However, at present there is a lack of suitable algorithms for calculating exam quality.
Methods: In the spring of 2014, the students' dean commissioned the „core group" for curricular improvement at the University Medical Center in Rostock to revise the criteria for the allocation of performance-based funds for teaching. In a first approach, we developed an algorithm that was based on the results of the most common type of exam in medical education, multiple choice tests. It included item difficulty and discrimination, reliability as well as the distribution of grades achieved.
Results: This algorithm quantitatively describes exam quality of multiple choice exams. However, it can also be applied to exams involving short assay questions and the OSCE. It thus allows for the quantitation of exam quality in the various subjects and – in analogy to impact factors and third party grants – a ranking among faculty.
Conclusion: Our algorithm can be applied to all test formats in which item difficulty, the discriminatory power of the individual items, reliability of the exam and the distribution of grades are measured. Even though the content validity of an exam is not considered here, we believe that our algorithm is suitable as a general basis for performance-based allocation of funds.

Keywords: exam quality, item difficulty, discrimination, reliability, performance based allocation of funds in teaching

Timo Kirschstein[1]
Alexander Wolters[1]
Jan-Hendrik Lenz[1]
Susanne Fröhlich[1]
Oliver Hakenberg[1]
Günther Kundt[2]
Martin Darmüntzel[3]
Michael Hecker[4]
Attila Altiner[3]
Brigitte Müller-Hilke[5]

1 Universitätsmedizin Rostock, "core group" zur Verbesserung der Lehre, Rostock, Deutschland

2 Universitätsmedizin Rostock, Institut für Biostatistik und Informatik in Medizin und Alternsforschung, Rostock, Deutschland

3 Universitätsmedizin Rostock, Studiendekanat, Rostock, Deutschland

4 Universitätsmedizin Rostock, Klinik und Poliklinik für Neurologie, Zentrum für Nervenheilkunde, Rostock, Deutschland

5 Universitätsmedizin Rostock, Institut für Immunologie, Rostock, Deutschland

Introduction

„Assessment drives learning". For the last 30 years, it has amply been analyzed and documented that we guide the learning styles and the academic performance of our students by the way we assess their knowledge [1], [2], [3], [4], [5]. In 2002, the amendment of the German Medical Licensing Act led to graded assessments in all clinical subjects and in an increasing number of interdisciplinary areas [http://www.gesetze-im-internet.de/_appro_2002/BJNR240500002.html]. As a general rule, these graded assessments are based on multiple choice (MC) tests.

This increase in graded assessments not only posed a logistic challenge for the faculties, but also offered the possibility to guide the students' learning behavior and to create the conditions for improved performance in the second state exam. The latter though required that the faculty specific exams are of high quality. To help ensure this quality, the German Society for Medical Education (GMA) together with the German Association of Medical Faculties (MFT) published recommendations for the administration of high-quality assessments [6], [7]. These

recommendations also provide quantifiable parameters like item difficulty and discrimination as well as the reliability of the exam as a whole. In general, the quantification of exam quality should be objective, reliable and valid. While objectivity and reliability can readily be quantified, validity can at best be estimated.

To meet the logistic requirements for the many written exams, the medical faculty of Rostock in 2009 implemented an electronic item management system, the use of which is voluntary yet accepted by almost all clinical departments. Ever since there is transparency on the results of all exams and those responsible for the exams obtain detailed feed-back on passing-scores, distribution of scores and grades achieved, item difficulty and discriminatory power of each item. Nonetheless, little has changed for the faculty wide assessments and not meeting the quality standards did not necessarily lead to noticeable efforts in improving MC exam quality. In order to direct the faculty's attention towards higher exam quality, we here decided to use exam quality as a criterion for calculating performance-based allocation of funds. However, in order to be accepted by the faculty and to lead to the desired effects, this calculation needed to be reproducible and transparent [8], [9]. Against this background, we here designed an algorithm to quantify exam quality as a basis on which to allocate performance based funds.

Methods

In the spring of 2014, the students' dean commissioned the „core group" for curricular improvement at the University Medical Center in Rostock to revise the criteria for the allocation of performance-based funds for teaching. As a first step towards the integration of exam quality, we assessed already published parameters for high quality exams like item difficulty, discrimination and reliability [10], [11], [12]. However, to additionally meet the observed asymmetry in the grading of some departments, we also calculated any deviation from the Gaussian distribution. Based on the results of all exams written in the clinical subjects taught in the summer term of 2014, we developed an algorithm that would include all four parameters equally and would allow for a ranking of the results. The basis for these calculations was a matrix showing for all students which item was answered correctly (1) or incorrectly (0) and what score was reached, respectively. These matrices were either generated out of the electronic item management system or were compiled manually. Even though type A/5 options is the most common type of items used in our written exams, some departments within the faculty also use short assay questions and our fifth year is required to sit an OSCE.

In a first step, we calculated the proportion of items per exam or stations per OSCE that featured both, an item difficulty between 0.40 and 0.85 and a part-whole-corrected discrimination characterized by a Pearson correlation coefficient (r) of at least 0.2. Item difficulty was here defined as the percentages of students who had correctly answered an MC question of Type A or the mean scores of short assay questions or of OSCE stations, respectively. Chi-square tests to evaluate the distribution of grades achieved as well as Cronbach's α were calculated in Excel. Subsequent correlation analyses performed with GraphPad InStat (Version 3) yielded Spearman-Rank correlation coefficients (r) and the corresponding 95%-confidence intervals (CI).

Results

We here used item difficulty, discrimination, reliability and grade distribution as objectively measurable quality criteria for written tests. These four parameters were reduced to three by defining a proportion of "good" questions that showed both an item difficulty between 0.4 and 0.85 and a part-whole-corrected Pearson correlation coefficient (r) of at least 0.2. Parameters two and three were reliability (described as Cronbach's α) and the distribution of the grades (described as the P value resulting from a chi-squared distribution test).

The proportion of "good" questions ranged per definition between 0 and 1. A value of 1 resulted if all questions of an exam perfectly fulfilled both criteria, namely item difficulty and discrimination. Cronbach's α can theoretically be negative however, evaluating written exams usually results in values between 0 and 1. As with the first parameter, "good" item, higher values represent better results. Chi-squared distribution of grades measures any deviation from the Gaussian distribution and resulting P-values smaller than 0.05 deny Gaussian distribution. In fact, the smaller the P-value, the more skewed the distribution of grades was. Evaluating all results obtained in the summer term in 2014 resulted in P-values between $4{,}6x10^{-158}$ and 0.99, respectively. Figure 1 presents two extreme grade distributions.

In order to weigh all three parameters equally, the resulting values describing the proportion of "good" questions, reliability (Cronbach's α) and the distribution of grades (P-value resulting from the chi-squared distribution test) were transformed onto a scale between 0 and 1 with the highest values being 1 and the lowest being 0. Subsequently, the transformed values describing the three parameters were added up and the results were ranked (see Table 1). Coefficients resulting from the ranks' correlation of the proportion of "good" items, reliability and distribution of grades were 0.660, 0.1229 and 0.1225, respectively.

Table 1 summarizes the quality of 19 exams that were written at the University Medical Center of Rostock in the summer of 2014 and that were evaluated using the electronic item management system. These exams were MC featuring the item type A, only. Introducing exam quality as another criteria for the performance-based allocation of funds led to the manual compilation of matrices for those exams that were not yet managed via the electronic system so that item difficulty, discrimina-

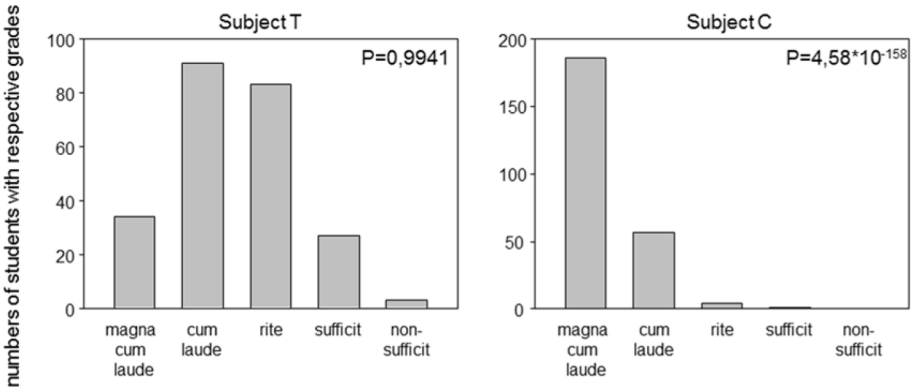

Figure 1: Extreme distributions of grades. Bar diagrams represent a Gaussian distribution of grades achieved for subjet T and a „ceiling effect" for subject C (letters denote the same subjects as in Table 1).

tion, reliability and distribution of grades could eventually be evaluated for all exams. Thus, our algorithm was applied not only to exams including short assay questions but also to the OSCE.

Discussion

Our algorithm presented here for evaluating exam and OSCE quality represents on the one hand internal consistencies – in the form of corrected discrimination and reliability – and on the other hand test results – in the form of item difficulty and grade distribution. Item difficulty, discrimination and reliability are already accepted as quality criteria in the medical literature [7], [13], [14], albeit we keep the lower limit of 0.4 for very difficult and the upper limit of 0.85 for very easy tasks also in short-assay tests and the OSCE. The lower limit could be redefined in non-MC test formats, however, one has to take into consideration that a further lowering might impair discriminatory power. Likewise, the upper limit is debated in the sense that exams should also include questions that can be answered by each and every student. Here, each faculty needs to issue individual recommendations as to the maximum proportion of "easy" questions.

The additional inclusion of grade distribution in our criteria is due to the observation that some departments consistently do not exploit the complete spectrum of possible grades (see Figure 1). The resulting skewedness precludes internal differentiation and is in our opinion not suitable to support the learning behaviour of students [5]. With the algorithm presented here, we aim at receiving a normal distribution, in which we intentionally do not declare the middle grade ("rite") as the mean, but allow for the individually calculated mean for each exam. We decided to assess the distribution of grades instead of scores achieved for two reasons,

1. there is no uniform maximum score in the written tests and
2. the scores achieved can potentially be normally distributed even if half of the students did not pass the exam.

Moreover, our algorithm is based on the P value which indicates the probability of deviation from the normal distribution, rather than the degree of deviation per se. We have decided to do so, because the P value and the 5% significance level are omnipresent and easy to understand. It remains to be considered that both the P value and the degree of deviation from the normal distribution depend on the sample size "N". This, however, is not critical as long as student cohorts are of approximately same sizes and the number of participants in the exams per year are comparable. In the algorithm presented here, the proportion of „good" questions correlated with the reliability. This could be partly due to a redundancy of the criteria assessed, but it is also likely that a dedicated examiner would produce not only questions with high discriminatory power, but also will take into account the distributions of item difficulty and include more question items, which in turn will raise reliability.

The algorithm presented here is a potential measure of good-quality exams. It can be applied for all test formats in which item difficulty, discrimination and grade distribution is recorded, and can therefore be used directly as a basis for performance-based funding. Once a transparent scoring system is established, item difficulty, discrimination and grade distribution can be calculated, even if the test formats are composite and performance is based on protocols or log books.

In contrast to the students' evaluation, which is most commonly used for the allocation of performance-based funding [9], the algorithm presented here has the advantage that exam quality is independent of the popularity of the subject. Nevertheless, this instrument also carries potential drawbacks: Subjects with high-quality written exams could insist on their reliable, but not necessarily valid test formats and thus prevent innovative changes. Here, the faculty could countersteer by not only assessing exam quality for performance-based funding, but also innovative teaching and learning formats. At the Rostock Medical School, performance-based funding for teaching consists of three criteria: exam quality, student evaluation and elective courses. Participation at the OSCE, an interdisciplinary event, is represented as an elective course where the quality of each station is used exclusively to guide the participating departments.

Table 1: Algorithm for the calculation of exam quality

Subject	relative frequency of „good“ items	relative frequency transformed	reliability (Cronbach's α)	reliability transformed	distribution of grades chi²-test P-value	distribution of grades transformed	Σ:“good“ items + reliability + distribution of grades	rank
I	0.59	0.69	0.90	1,00	0.39	0.39	2.079	**1**
D	0.54	0.62	0.75	0,79	0.47	0.48	1.888	**2**
O	0.86	1.00	0.82	0,89	$2.3x10^{-05}$	0.00	1.886	**3**
P	0.19	0.22	0.80	0,85	0.43	0.43	1.502	**4**
F	0.13	0.15	0.47	0,36	0.90	0.91	1.426	**5**
H	0.03	0.04	0.64	0,62	0.59	0.60	1.253	**6**
T	0.10	0.11	0.30	0,11	0.99	1.00	1.223	**7**
Q	0.20	0.23	0.59	0,54	0.38	0.38	1.142	**8**
E	0.25	0.29	0.79	0,84	$1.8x10^{-03}$	0.00	1.132	**9**
B	0.22	0.25	0.52	0,44	0.34	0.34	1.035	**10**
N	0.15	0.18	0.41	0,27	0.46	0.46	0.910	**11**
R	0.10	0.12	0.54	0,46	0.15	0.15	0.730	**12**
K	0.19	0.22	0.45	0,33	$1.6x10^{-08}$	0.00	0.553	**13**
S	0.10	0.12	0.44	0,32	0.04	0.04	0.482	**14**
A	0.21	0.25	0.32	0,14	0.049	0.05	0.437	**15**
G	0.00	0.00	0.50	0,41	0.00	0.00	0.409	**16**
L	0.08	0.09	0.36	0,21	0.04	0.04	0.340	**17**
M	0.00	0.00	0.23	0,00	0.30	0.30	0.301	**18**
C	0.00	0.00	0.27	0,06	$4.6x10^{-158}$	0.00	0.057	**19**
r = 0.66; CI= 0.28 – 0.86; P=0.0021			r = 0.12; CI= -0.36 – 0.56; P=0.6163					
r = 0.12; CI= 0.36 – 0.56; P=0.6174								

Whether indeed exam quality can be improved by the allocation of funds will only transpire after testing and evaluating this control instrument for several years. However, we are optimistic that the modified funding for teaching can at least draw more attention towards exams. A first objective, namely that those subjects who did not use the electronic exam management system before, now analyze their exams on qualitative criteria, has been achieved. The transparency of the applied criteria – item difficulty, discrimination and grades – offers the opportunity that teachers intensify reflection on their exams and seek to improve test quality [8], [15]. The algorithm presented here offers several possibilities for adjustment, among them reliability, that can most easily be influenced by the number of question items. In the past, a lack of discriminatory power has sporadically been used to review the distractors and to check the conformity of exam and course content. Ideally, the algorithm presented here will not only help to improve the quality of the individual question items, but will motivate the faculty members to question the validity of their tests, a parameter that cannot be assessed with our instrument. Ultimately, it remains to be seen whether and how an improved test quality will impact on the student evaluation on the one hand and on the performance in the second written state exam on the other hand.

Competing interests

The authors declare that they have no competing interests.

References

1. Biggs J. Enhancing teaching through constructive alignment. High Educ. 1996;32:347-364. DOI: 10.1007/BF00138871
2. Shumway JM, Harden RM; Association for Medical Education in E. AMEE Guide No. 25: The assessment of learning outcomes for the competent and reflective physician. Med Teach. 2003;25(6):569-584. DOI: 10.1080/0142159032000151907
3. chuwirth L, van der Vleuten C. Merging views on assessment. Med Educ. 2004;38(12):1208-1210. DOI: 10.1111/j.1365-2929.2004.02055.x
4. Wormald BW, Schoeman S, Somasunderam A, Penn M. Assessment drives learning: an unavoidable truth? Anat Sci Educ. 2009;2(5):199-204. DOI: 10.1002/ase.102
5. 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. DOI: 10.3109/0142159X.2011.551559
6. Gesellschaft für Medizinische Ausbildung, Kompetenzzentrum Prüfungen Baden-Württemberg, Fischer MR. Leitlinie für Fakultäts-interne Leistungsnachweise während des Medizinstudiums: Ein Positionspapier des GMA-Ausschusses Prüfungen und des Kompetenzzentrums Prüfungen Baden-Württemberg. GMS Z Med Ausbild. 2008;25(1):Doc74. Zugänglich unter/available from: http://www.egms.de/static/de/journals/zma/2008-25/zma000558.shtml
7. Jünger J, Just I. Empfehlungen der Gesellschaft für Medizinische Ausbildung und des Medizinischen Fakultätentags für fakultätsinterne Leistungsnachweise während des Studiums der Human-, Zahn- und Tiermedizin. GMS Z Med Ausbild. 2014;31(3):Doc34. DOI: 10.3205/zma000926
8. Kreysing M. Forschungsförderung mittels leistungsorientierter Mittelvergabe. Z Hochschulentw. 2008;3:19-28.
9. Müller-Hilke B. "Ruhm und Ehre" oder LOM für Lehre? - eine qualitative Analyse von Anreizverfahren für gute Lehre an Medizinischen Fakultäten in Deutschland. GMS Z Med Ausbild. 2010;27(3):Doc43. DOI: 10.3205/zma000680
10. Möltner A, Schellberg D, Jünger J. Grundlegende quantitative Analysen medizinischer Prüfungen. GMS Z Med Ausbild. 2006;23(3):Doc53. Zugänglich unter/available from: http://www.egms.de/static/de/journals/zma/2006-23/zma000272.shtml
11. WFME, AWSE. WFME Global Standards for Quality improvement in Medical Education European Specifications. Copenhagen: University of Copenhagen, MEDLINE Quality Assurance Task Force; 2007.
12. WHO, WFME. Guidelines for Accreditation of Basic Medical Education. Geneva, Copenhagen: WHO; 2005.
13. Norcini J, Anderson B, Bollela V, Burch V, Costa MJ, Duvivier R, Galbraith R, Hays R, Kent A, Perrott V, Roberts T. Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 Conference. Med Teach. 2011;33(3):206-214. DOI: 10.3109/0142159X.2011.551559
14. Schuwirth LW, Van der Vleuten CP. Programmatic assessment: From assessment of learning to assessment for learning. Med Teach. 2011;33(6):478-485. DOI: 10.3109/0142159X.2011.565828
15. Müller-Hilke B. Leistungsorientierte Mittelvergabe. Mehr Geld zur Lehre lenken. Dtsch Arztebl. 2013;110:A-2418/B-2131/C-2053.

Corresponding author:
Brigitte Müller-Hilke
Universitätsmedizin Rostock, Institut für Immunologie, Schillingallee 69, D-18057 Rostock, Deutschland, Tel.: +49 (0)381/4945-883, Fax: +49 (0)381/4945-882
brigitte.mueller-hilke@med.uni-rostock.de

Please cite as
Kirschstein T, Wolters A, Lenz JH, Fröhlich S, Hakenberg O, Kundt G, Darmüntzel M, Hecker M, Altiner A, Müller-Hilke B. An algorithm for calculating exam quality as a basis for performance-based allocation of funds at medical schools. GMS J Med Educ. 2016;33(3):Doc44. DOI: 10.3205/zma001043, URN: urn:nbn:de:0183-zma0010437

This article is freely available from
http://www.egms.de/en/journals/zma/2016-33/zma001043.shtml

Received: *2015-02-21*
Revised: *2016-02-02*
Accepted: *2016-03-04*
Published: *2016-05-17*

OPEN ACCESS

This is the translated (German) version.
The original (English) version starts at p. 1.

Ein Algorithmus zur Berechnung von Klausurqualität als Bemessungsgrundlage für LOM-Lehre

Timo Kirschstein[1]
Alexander Wolters[1]
Jan-Hendrik Lenz[1]
Susanne Fröhlich[1]
Oliver Hakenberg[1]
Günther Kundt[2]
Martin Darmüntzel[3]
Michael Hecker[4]
Attila Altiner[3]
Brigitte Müller-Hilke[5]

1 Universitätsmedizin Rostock, "core group" zur Verbesserung der Lehre, Rostock, Deutschland

2 Universitätsmedizin Rostock, Institut für Biostatistik und Informatik in Medizin und Alternsforschung, Rostock, Deutschland

3 Universitätsmedizin Rostock, Studiendekanat, Rostock, Deutschland

4 Universitätsmedizin Rostock, Klinik und Poliklinik für Neurologie, Zentrum für Nervenheilkunde, Rostock, Deutschland

5 Universitätsmedizin Rostock, Institut für Immunologie, Rostock, Deutschland

Zusammenfassung

Zielsetzung: Mit der Novellierung der ÄAppO im Jahr 2002 und der Einführung benoteter Leistungsnachweise hat die Bedeutung schriftlicher Prüfungen im klinischen Abschnitt des Medizinstudiums deutlich zugenommen. Allerdings werden die Mindestanforderungen an die Qualität von Prüfungen mitunter nur schwer erreicht. Mit der Aufnahme von Prüfungsqualität in den Kriterienkatalog zur Vergabe von LOM (Leistungsorientierter Mittelvergabe) könnte die Aufmerksamkeit der Lehrenden gelenkt und allein dadurch die Qualität der Prüfungen verbessert und verstetigt werden. Derzeit mangelt es jedoch an geeigneten Bemessungsgrundlagen.
Methoden: Im Frühjahr 2014 wurde die „core group" zur Verbesserung der Lehre an der Universitätsmedizin Rostock vom Studiendekan beauftragt, den der Verteilung von LOM Lehre zugrunde liegenden Kriterienkatalog zu überarbeiten. In diesem Zusammenhang wurde zunächst anhand von multiple choice-Klausurergebnissen ein Algorithmus entwickelt, der auf Aufgabenschwierigkeit, Trennschärfe, Reliabilität und Notenspiegel basiert und damit die Qualität der häufigsten Prüfungsform im Studium der Humanmedizin quantitativ abbildet.
Ergebnisse: Dieser Algorithmus wurde anschließend auch auf Klausuren mit offenen Fragen sowie auf den OSCE übertragen. Mit seiner Hilfe lässt sich die Prüfungsqualität in den einzelnen Fächern berechnen und – vergleichbar mit Impaktpunkten und Drittmitteleinwerbungen - in eine intrafakultäre Rangfolge überführen.
Schlussfolgerung: Dieser Algorithmus ist auf alle Prüfungsformate anwendbar, bei denen Aufgabenschwierigkeit, Trennschärfe, Reliabilität und Notenspiegel erfasst werden. Auch wenn eine weitere wichtige Kenngröße, nämlich die Validität einer Prüfung hier nicht berücksichtigt wird, so ist der vorgestellte Algorithmus als Bemessungsgrundlage für LOM durchaus geeignet.

Schlüsselwörter: Prüfungsqualität, Aufgabenschwierigkeit, Trennschärfe, Reliabilität, LOM Lehre

Einleitung

„Assessment drives learning". Seit etwa 30 Jahren wird ausführlich untersucht und belegt, dass die Art und Weise wie wir prüfen, das Lernverhalten und den Lernerfolg der Studierenden maßgeblich beeinflusst [1], [2], [3], [4], [5]. Mit der Novellierung der ÄAppO 2002 wurde für alle klinischen Fächer und eine steigende Anzahl von Querschnittsbereichen ein benoteter Leistungsnachweis eingeführt, der in aller Regel über eine schriftliche multiple choice (MC) Klausur erbracht wird http://www.gesetze-im-internet.de/_appro_2002/BJNR240500002.html]. Damit ergab sich für die Medizinischen Fakultäten zwar eine neue logistische Herausforderung - aber auch die Chance, über diese Klausuren das Lernverhalten der Studierenden zu steuern und im Idealfall die Voraussetzung für ein gutes Abschneiden im 2. Staatsexamen zu schaffen. Letzteres erfordert jedoch, dass die fakultätsinternen Prüfungen einem hohen Qualitätsstandard folgen. Hier erfahren die Fakultäten Unterstützung in der Form von Leitlinien, die die Qualitätskriterien von schriftlichen und mündlichen Prüfungen zusammenfassen [6], [7]. Zu diesen Qualitäts-

kriterien gehören gut abbild- und berechenbare Parameter wie Schwierigkeit und Trennschärfe einzelner Aufgaben sowie die Reliabilität einer gesamten Prüfung. Die Erhebung der Qualitätskriterien für die Erfassung von Prüfungsqualität sollte objektivierbar, reliabel und valide sein. Während sich Objektivität und Reliabilität quantifizieren lassen, kann die Validität allenfalls geschätzt werden.
Um der logistischen Herausforderung durch die Fülle der schriftlichen Klausuren zu begegnen, hat die Universitätsmedizin Rostock (UMR) 2009 ein elektronisches Prüfungsmanagement eingeführt, das seit 2011 als freiwilliges Angebot fast flächendeckend im klinischen Studienabschnitt eingesetzt wird. Seitdem erlangt das Studiendekanat Einblick in alle Prüfungsergebnisse, während die Prüfungsverantwortlichen eine detaillierte Rückkopplung über Bestehensgrenzen, Notenspiegel, Aufgabenschwierigkeiten, Antworthäufigkeiten und Trennschärfen erhalten. Dennoch hat sich am Prüfungsverhalten der Einrichtungen wenig verändert und das Nichterreichen der angestrebten Qualitätsstandards mündete nicht zwangsläufig in einer wahrnehmbaren Bestrebung, schriftliche MC-Prüfungen qualitativ zu verbessern. Als aufmerksamkeitssteigernde Maßnahme soll deswegen die leistungsorientierte Mittelvergabe (LOM) genutzt werden. Um jedoch Akzeptanz innerhalb der Fakultät und messbare Verhaltensänderungen zu bewirken, müssen die Kriterien für die LOM nachvollziehbar und transparent sein [8], [9]. Vor diesem Hintergrund wurde ein Algorithmus zur Quantifizierung von Prüfungsqualität entwickelt, der als Bemessungsgrundlage für LOM Lehre herangezogen werden kann.

Methoden

Im Frühjahr 2014 wurde die „core group“ zur Verbesserung der Lehre an der UMR vom Studiendekan beauftragt, den der Verteilung von LOM Lehre zugrunde liegenden Kriterienkatalog zu überarbeiten. Um die Prüfungsqualität als Kenngröße integrieren zu können, wurden zunächst bereits publizierte Kriterien qualitativ hochwertiger Klausuren wie Aufgabenschwierigkeit, Trennschärfe und Reliabilität der einzelnen Fragenitems berücksichtigt [10], [11], [12]. Um einer möglichen Schiefe bei der Notenvergabe zu begegnen, wurde zusätzlich die Normalverteilung der Noten als weiteres Kriterium zur Abschätzung von Prüfungsqualität hinzugezogen. Anhand der Ergebnisse aller im Sommersemester 2014 an der UMR geschriebenen Klausuren im klinischen Studienabschnitt wurde ein Algorithmus entwickelt, mit dem alle einzuschließenden Kriterien gleichwertig quantifiziert und die Summen in eine Rangfolge transformiert wurden. Grundlage für die nachfolgenden Berechnungen waren aus den jeweiligen Prüfungen generierte Matrizes, die für jeden Klausurteilnehmer die Information enthalten, welche Frage richtig (1) oder falsch (0) beantwortet bzw. wie viele Punkte bei einer bestimmten Aufgabe erzielt wurden. Diese Matrizes werden bei uns entweder aus dem elektronischen Prüfungsmanagement heraus generiert oder händisch erstellt. Der Hauptfragentyp bei unseren Klausuren ist Typ A mit 5 Antwortoptionen, wobei es auch Fächer mit short essay questions und im 10. Semester einen OSCE gibt.
Zunächst wurde der Anteil der Fragen bzw. Prüfstationen ermittelt, die sowohl eine Aufgabenschwierigkeit zwischen 0,40 und 0,85 als auch eine part-whole-korrigierte Trennschärfe von r=0,2 (Pearson Korrelationskoeffizient) aufwiesen. Aufgabenschwierigkeit war dabei definiert als der Anteil der Studierenden, die eine MC-Frage vom Typ A zutreffend beantwortet haben bzw. die mittlere erreichte Punktzahl bei short essay questions und OSCE-Stationen. Der Chi-Quadrat-Test zur Berechnung der Verteilung der Noten sowie die Berechnung des jeweiligen Cronbach's α wurden in Excel durchgeführt. Für die nachfolgenden Korrelationsanalysen wurden der Spearmans-Rangkorrelationskoeffizient (r) sowie das dazugehörige 95%-Konfidenzintervall (CI) in GraphPad InStat (Version 3) berechnet.

Ergebnisse

Als objektiv messbare Qualitätskriterien für schriftliche Prüfungen wurden Aufgabenschwierigkeit, Trennschärfe, Reliabilität und Notenverteilung für die Berechnung von Qualität herangezogen. Diese vier Parameter wurden auf drei Messgrößen reduziert, nämlich ein Anteil an „guten“ Fragen (Aufgabenschwierigkeit zwischen 0,4 und 0,85 bei gleichzeitiger (korrigierter) Trennschärfe von r=0,2), die Reliabilität (beschrieben als Cronbach's α) und die Notenverteilung (beschrieben als P-Wert des Chi-Quadrat-Verteilungstests).
Als erste Messgröße wurde der Anteil „guter“ Fragen herangezogen, der definitionsgemäß im Intervall zwischen 0 und 1 liegt. Erfüllen sämtliche Fragen einer Klausur die beiden Kriterien (Aufgabenschwierigkeit und Trennschärfe), würde ein Wert von 1 resultieren. Cronbach's α kann theoretisch auch negative Werte annehmen, liegt aber bei der Auswertung von schriftlichen Prüfungen in der Regel zwischen 0 und 1. Auch hier stehen höhere Werte für eine bessere Qualität. Der Chi-Quadrat-Verteilungstest misst die Abweichung der Notenverteilung von der Normalverteilung. Ein P-Wert kleiner als 0,05 spricht gegen eine Normalverteilung und je kleiner der P-Wert wird, desto größer ist die Schieflage der Notenverteilung. Bei dem von uns ausgewerteten Klausursemester lagen die resultierenden P-Werte zwischen $4{,}6 \times 10^{-158}$ und 0,99. Abbildung 1 zeigt exemplarisch zwei extreme Notenverteilungen.
Um allen drei Messgrößen bei der Qualitätsberechnung das gleiche Gewicht zu verleihen, wurden die resultierenden Werte – Anteil „guter“ Fragen, Cronbach's α, P-Wert aus dem Chi2-Test - auf eine Skala zwischen 0 und 1 transformiert. Abschließend wurden für jedes Fach die Summe dieser drei transformierten Werte gebildet und eine Rangfolge erstellt (siehe Tabelle 1). Die aus den Korrelationen der Messwerte für den Anteil „guter“ Fragen, die Reliabilität und die Notenverteilung resultieren-

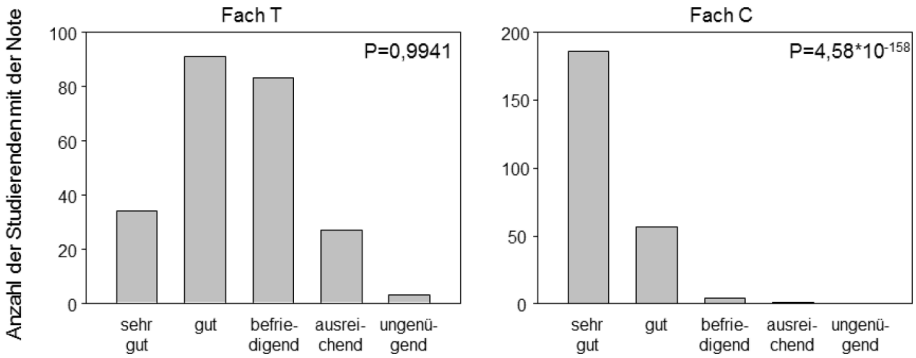

Abbildung 1: Extreme der Notenverteilung. Die Balkendiagramme verdeutlichen eine Normalverteilung der Noten im Fach T und einen „Deckeneffekt" im Fach C (Bezeichnungen der Fächer entsprechen denjenigen in Tabelle 1).

den Rangkorrelationskoeffizienten betragen 0,660, 0,1229 bzw. 0,1225 (siehe Tabelle 1).

Tabelle 1 bildet die berechnete Qualität für 19 Klausuren ab, die an der UMR im Sommersemester in klinischen Abschnitt des Medizinstudiums geschrieben und mir dem elektronischen Prüfungssystem erfasst wurden. Dabei handelt es sich ausschließlich um Klausuren mit MC-Fragen vom Typ A. Mit der Aufnahme von Klausurqualität in den Kriterienkatalog für LOM-Lehre wurden dann auch Klausuren, die nicht elektronisch erfasst wurden, händisch in Matrizes übertragen, so dass Schweregrad, Trennschärfe, Reliabilität und Notenverteilung berechnet werden konnten. Darunter waren auch Klausuren mit short essay questions. Für den OSCE, der an der UMR zu Beginn des 10. Semesters stattfindet, werden gleichfalls Schweregrad, Trennschärfe, Reliabilität und Notenspiegel berechnet.

Diskussion

Unser hier vorgestellter Algorithmus zur Bewertung von Klausur- und OSCE-Qualität bildet zum einen die interne Konsistenz – in der Form von korrigierter Trennschärfe und Reliabilität – und zum anderen das Prüfungsergebnis – in der Form von Aufgabenschwierigkeit und Notenspiegel – ab. Aufgabenschwierigkeit, Trennschärfe und Reliabilität sind in der medizinischen Literatur bereits als Qualitätskriterien akzeptiert [7], [13], [14], wobei wir an dem unteren Grenzwert von 0,4 für sehr schwere und an dem oberen von 0,85 für sehr leichte Aufgaben auch bei short essay-Fragen und bei OSCE-Stationen festhalten. Der untere Wert könnte bei nicht-MC-Formaten auch anders festgelegt werden, wobei bei einem weiteren Absenken darauf geachtet werden sollte, dass die Aufgaben trennscharf bleiben. Auch der obere Wert wird regelmäßig vor dem Hintergrund diskutiert, dass Klausuren auch solche Fragen enthalten sollen, die jeder Student beantworten kann. Hier könnte jede Fakultät ihre eigenen Empfehlungen aussprechen, wie hoch der Anteil dieser sehr leichten Fragen maximal sein sollte.

Die zusätzliche Aufnahme des Notenspiegels in unseren Kriterienkatalog ist der Beobachtung geschuldet, dass einige Fächer das Notenspektrum konsequent nicht ausnutzen (siehe Abbildung 1). Die resultierende Schieflage verhindert die Binnendifferenzierung und ist aus unserer Sicht nicht geeignet, das Lernverhalten der Studierenden zu unterstützen [5]. Mit dem hier vorgestellten Algorithmus streben wir eine Normalverteilung an, wobei wir den Mittelwert bewusst nicht bei „befriedigend" festlegen, sondern den für jede Klausur individuell berechneten Mittelwert zulassen. Statt der erreichten Punkte legen wir die erreichten Noten für die Berechnung der Normalverteilung zugrunde, weil es einerseits keine einheitliche maximale Punktzahl in unseren Prüfungen gibt und andererseits die erreichten Punkte auch dann noch normalverteilt sein könnten, wenn die Hälfte der Studierenden die Prüfung nicht bestanden hat. Außerdem geht in unseren Algorithmus statt des Ausmaßes der Abweichung von einer Normalverteilung der P-Wert, der die Wahrscheinlichkeit der Abweichung beschreibt, ein. Für diesen Wert haben wir uns entschieden, weil der P-Wert und das 5%ige Signifikanzniveau omnipräsent und leicht nachvollziehbar sind. Dabei bleibt zu berücksichtigen, dass sowohl der P-Wert als auch der das Ausmaß der Abweichung von einer Normalverteilung beschreibende statistische Wert von „N" abhängig sind. Das stellt jedoch dann kein Problem, wenn die klinischen Jahrgänge in etwa gleich groß und damit die Teilnehmerzahlen an den Klausuren pro Jahr vergleichbar sind.

Bei dem hier vorgestellten Algorithmus korreliert der Anteil „guter" Fragen mit der Reliabilität. Das könnte zum einen auf eine Redundanz der Messkriterien zurückzuführen sein, zum anderen könnte es aber auch kausale Gründe geben, wonach ein engagierter Prüfer möglicherweise nicht nur Fragen mit hoher Trennschärfe konzipiert, sondern auch die Verteilung der Aufgabenschwierigkeiten berücksichtigt und mehr Fragenitems inkludiert, wodurch wiederum die Reliabilität angehoben wird.

Der hier vorgestellte Algorithmus ist ein mögliches Messinstrument für qualitativ gute Prüfungen. Er ist auf alle Prüfungsformate, bei denen Aufgabenschwierigkeit, Trennschärfe und Notenspiegel ermittelt werden übertragbar und kann somit direkt als Bemessungsgrundlage für LOM eingesetzt werden. Auch bei zusammengesetzten Prüfungsformaten, bei denen Protokolle oder Log-Bücher in den Leistungsnachweis eingehen, sind die Berechnung von Aufgabenschwierigkeit, Trennschärfe, Reliabilität und Notenspiegel denkbar, sobald ein nachvollziehbares Punktesystem zugrunde gelegt wird.

Tabelle 1: Algorithmus zur Berechnung von Prüfungsqualität

Fach	Relative Häufigkeit „guter“ Fragen	Relative Häufigkeit transformiert	Reliabilität (Cronbach's α)	Reliabilität transformiert	Notenverteilung chi^2-Test P-Wert	Notenverteilung transformiert	Σ:gute Fragen + Reliabilität + Notenverteilung	**Rang**
I	0,59	0,69	0,90	1,00	0,39	0,39	2,079	**1**
D	0,54	0,62	0,75	0,79	0,47	0,48	1,888	**2**
O	0,86	1,00	0,82	0,89	$2,3x10^{-05}$	0,00	1,886	**3**
P	0,19	0,22	0,80	0,85	0,43	0,43	1,502	**4**
F	0,13	0,15	0,47	0,36	0,90	0,91	1,426	**5**
H	0,03	0,04	0,64	0,62	0,59	0,60	1,253	**6**
T	0,10	0,11	0,30	0,11	0,99	1,00	1,223	**7**
Q	0,20	0,23	0,59	0,54	0,38	0,38	1,142	**8**
E	0,25	0,29	0,79	0,84	$1,8x10^{-03}$	0,00	1,132	**9**
B	0,22	0,25	0,52	0,44	0,34	0,34	1,035	**10**
N	0,15	0,18	0,41	0,27	0,46	0,46	0,910	**11**
R	0,10	0,12	0,54	0,46	0,15	0,15	0,730	**12**
K	0,19	0,22	0,45	0,33	$1,6x10^{-08}$	0,00	0,553	**13**
S	0,10	0,12	0,44	0,32	0,04	0,04	0,482	**14**
A	0,21	0,25	0,32	0,14	0,049	0,05	0,437	**15**
G	0,00	0,00	0,50	0,41	0,00	0,00	0,409	**16**
L	0,08	0,09	0,36	0,21	0,04	0,04	0,340	**17**
M	0,00	0,00	0,23	0,00	0,30	0,30	0,301	**18**
C	0,00	0,00	0,27	0,06	$4,6x10^{-158}$	0,00	0,057	**19**

r = 0,66; CI= 0,28 – 0,86; P=0,0021 — r = 0,12; CI= -0,36 - 0,56; P=0,6163

r = 0,12; CI= 0,36 - 0,56; P=0,6174

Gegenüber der studentischen Evaluation, die am häufigsten für die Vergabe von LOM-Lehre herangezogen wird [9], bietet der hier vorgestellte Algorithmus den großen Vorteil, dass Prüfungsqualität unabhängig von der Beliebtheit eines Faches ist. Gleichwohl birgt dieses Instrument auch mögliche Nachteile: So könnten Fächer mit einer hohen Klausurqualität aufgrund der LOM auf ihrem reliablen, aber nicht zwangsläufig validen Prüfungsformat beharren und innovative Änderungen unterbinden. Hier könnte die Fakultät gegensteuern, indem nicht nur die Prüfungsqualität Eingang in die LOM-Berechnung findet, sondern z.B. auch innovative Lehr- und Lernformate. An der UMR setzt sich die LOM-Lehre aus den drei Kriterien Prüfungsqualität, studentische Evaluation und zusätzliches Lehrangebot zusammen. Die Beteiligung am OSCE, einer interdisziplinären Veranstaltung, bildet sich im zusätzlichen Lehrangebot ab und die Qualität der einzelnen Stationen wird ausschließlich genutzt, um prüfende Einrichtungen zu steuern.

Ob sich die Prüfungsqualität tatsächlich durch den Zufluss von Mitteln verbessern lässt, wird sich erst herausstellen, wenn dieses Steuerungsinstrument einige Jahre erprobt und evaluiert worden ist. Wir sind jedoch optimistisch, dass sich durch den veränderten Mittelfluss zumindest die Aufmerksamkeit in Richtung Prüfungen lenken lässt. Ein erstes Ziel, dass nämlich auch diejenigen Fächer, die das elektronische Prüfungsmanagement nicht nutzen, ihre Klausuren jetzt auf qualitative Kriterien analysieren, ist bereits erreicht. Die Transparenz der angelegten Kriterien – Schweregrad, Trennschärfe, Reliabilität und Notenspiegel – birgt darüber hinaus eine hohe Wahrscheinlichkeit, dass die Lehrenden die Reflexion über ihre Prüfungen intensivieren und eine verbesserte Prüfungsqualität anstreben [8], [15]. Der hier vorgestellte Algorithmus bietet dazu mehrere Stellschrauben, wovon die Reliabilität über die Anzahl der Fragenitems am leichtesten zu beeinflussen ist. Ein Mangel an Trennschärfe wird bereits vereinzelt genutzt, um die Distraktoren zu analysieren und die Übereinstimmung von Prüfungs- und Lehrinhalten zu kontrollieren. Im Idealfall wird sich also nicht nur die Qualität der einzelnen Fragenitems verbessern, sondern die Lehrverantwortlichen hinterfragen auch die Validität ihrer Prüfungen, die wir mit unserem Messinstrument direkt gar nicht erfassen können. Letztendlich bleibt auch abzuwarten, ob und wie sich eine verbesserte Prüfungsqualität auf die studentische Evaluation einerseits und das Abschneiden der Studierenden im zweiten schriftlichen Staatsexamen andererseits auswirken wird.

Interessenkonflikt

Die Autoren erklären, dass sie keinen Interessenkonflikt im Zusammenhang mit diesem Artikel haben.

Literatur

1. Biggs J. Enhancing teaching through constructive alignment. High Educ. 1996;32:347-364. DOI: 10.1007/BF00138871
2. Shumway JM, Harden RM; Association for Medical Education in E. AMEE Guide No. 25: The assessment of learning outcomes for the competent and reflective physician. Med Teach. 2003;25(6):569-584. DOI: 10.1080/0142159032000151907
3. chuwirth L, van der Vleuten C. Merging views on assessment. Med Educ. 2004;38(12):1208-1210. DOI: 10.1111/j.1365-2929.2004.02055.x
4. Wormald BW, Schoeman S, Somasunderam A, Penn M. Assessment drives learning: an unavoidable truth? Anat Sci Educ. 2009;2(5):199-204. DOI: 10.1002/ase.102
5. 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. DOI: 10.3109/0142159X.2011.551559
6. Gesellschaft für Medizinische Ausbildung, Kompetenzzentrum Prüfungen Baden-Württemberg, Fischer MR. Leitlinie für Fakultäts-interne Leistungsnachweise während des Medizinstudiums: Ein Positionspapier des GMA-Ausschusses Prüfungen und des Kompetenzzentrums Prüfungen Baden-Württemberg. GMS Z Med Ausbild. 2008;25(1):Doc74. Zugänglich unter/available from: http://www.egms.de/static/de/journals/zma/2008-25/zma000558.shtml
7. Jünger J, Just I. Empfehlungen der Gesellschaft für Medizinische Ausbildung und des Medizinischen Fakultätentags für fakultätsinterne Leistungsnachweise während des Studiums der Human-, Zahn- und Tiermedizin. GMS Z Med Ausbild. 2014;31(3):Doc34. DOI: 10.3205/zma000926
8. Kreysing M. Forschungsförderung mittels leistungsorientierter Mittelvergabe. Z Hochschulentw. 2008;3:19-28.
9. Müller-Hilke B. "Ruhm und Ehre" oder LOM für Lehre? - eine qualitative Analyse von Anreizverfahren für gute Lehre an Medizinischen Fakultäten in Deutschland. GMS Z Med Ausbild. 2010;27(3):Doc43. DOI: 10.3205/zma000680
10. Möltner A, Schellberg D, Jünger J. Grundlegende quantitative Analysen medizinischer Prüfungen. GMS Z Med Ausbild. 2006;23(3):Doc53. Zugänglich unter/available from: http://www.egms.de/static/de/journals/zma/2006-23/zma000272.shtml
11. WFME, AWSE. WFME Global Standards for Quality improvement in Medical Education European Specifications. Copenhagen: University of Copenhagen, MEDLINE Quality Assurance Task Force; 2007.
12. WHO, WFME. Guidelines for Accreditation of Basic Medical Education. Geneva, Copenhagen: WHO; 2005.
13. Norcini J, Anderson B, Bollela V, Burch V, Costa MJ, Duvivier R, Galbraith R, Hays R, Kent A, Perrott V, Roberts T. Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 Conference. Med Teach. 2011;33(3):206-214. DOI: 10.3109/0142159X.2011.551559
14. Schuwirth LW, Van der Vleuten CP. Programmatic assessment: From assessment of learning to assessment for learning. Med Teach. 2011;33(6):478-485. DOI: 10.3109/0142159X.2011.565828
15. Müller-Hilke B. Leistungsorientierte Mittelvergabe. Mehr Geld zur Lehre lenken. Dtsch Arztebl. 2013;110:A-2418/B-2131/C-2053.

Korrespondenzadresse:
Brigitte Müller-Hilke
Universitätsmedizin Rostock, Institut für Immunologie, Schillingallee 69, D-18057 Rostock, Deutschland, Tel.: +49 (0)381/4945-883, Fax: +49 (0)381/4945-882
brigitte.mueller-hilke@med.uni-rostock.de

Bitte zitieren als
Kirschstein T, Wolters A, Lenz JH, Fröhlich S, Hakenberg O, Kundt G, Darmüntzel M, Hecker M, Altiner A, Müller-Hilke B. An algorithm for calculating exam quality as a basis for performance-based allocation of funds at medical schools. GMS J Med Educ. 2016;33(3):Doc44. DOI: 10.3205/zma001043, URN: urn:nbn:de:0183-zma0010437

Artikel online frei zugänglich unter
http://www.egms.de/en/journals/zma/2016-33/zma001043.shtml

Eingereicht: *21.02.2015*
Überarbeitet: *02.02.2016*
Angenommen: *04.03.2016*
Veröffentlicht: *17.05.2016*