The Final Oral/Practical State Examination at Freiburg Medical Faculty in 2012 – Analysis of grading to test quality assurance

Abstract

Aim: The aim of this study is to analyze the grades given for the oral/practical part of the German State Examination at the Medical Faculty of Freiburg. We examined whether or not the grades given for the written and the oral/practical examinations correlated and if differences in grading between the Freiburg University Medical Center (UMC) and the other teaching hospitals could be found.
In order to improve the quality of the state examination, the medical school has been offering standardized training for examiners for several years. We evaluated whether or not trained and untrained examiners differed in their grading of the exam and how these differences have changed over time.
Methods: The results of the 2012 spring and fall exams were analyzed (N=315). The relevant data set was made available to us by the Baden-Württemberg Examination Office (*Landesprüfungsamt*). The data were analyzed by means of descriptive and inferential statistics.
Results: We observed a correlation of ρ=0.460** between the grades for the written and the oral/practical exams. The UMC and the teaching hospitals did not differ significantly in their grade distributions. Compared to untrained examiners, trained ones assigned the grade of "very good" less often. Furthermore, they displayed a significantly higher variance in the grades given (p=0.007, phi=0.165). This effect is stronger when concentrating specifically on those examiners who took part in the training less than a year before.
Conclusion: The results of this study suggest that the standardized training for examiners at the Medical Faculty of Freiburg is effective for quality assurance. As a consequence, more examiners should be motivated to take part in the training.

Keywords: German State Examination, oral and practical part of the state examination, training of examiners, grading

Angela Schickler[1]
Peter Brüstle[1]
Silke Biller[1]

1 Uni Freiburg, Kompetenzzentrum Lehrevaluation in der Medizin Baden-Württemberg, Sitz Freiburg, Freiburg, Deutschland

Background

The 2002 medical licensing regulations (*Ärztliche Approbationsordnung*) [1] led to fundamental changes in the German State Examination, which then consisted of two sections until the current changes in 2012. The first section (M1) was taken by students after the first two years of study; the second one (M2) following the entire course of study, including the final practical year. Both sections are divided into a written and an oral/practical part. Some of the changes in 2002 affected the oral/practical part of the M2: its score represents one-third of the student's final overall grade. The length of the test was extended to two full days and a practical section was explicitly introduced. A total of four to five examiners administer the examination to a maximum of four examinees [1]. It is required that the focus of this examination be on patient-oriented questions [1], [2], [3].

In general, the oral/practical M2 examination is a very extensive, professionally qualifying exam that places considerable demands on the examiners. In a 2011 resolution, the *Medizinische Fakultätentag* (MFT) made reference to the great burden placed on the medical schools. It raised concern that through this increased use of resources and the necessary expansion of the examiner pool to include people from outside the universities, a decrease in exam quality and unequal treatment of the candidates could possibly result [http://www.mft-online.de/files/200_omft_2011.pdf most recently verified on 5 Nov. 2013]. To train final-year medical students, the Medical Faculty of Freiburg cooperates with 15 teaching hospitals, which are also involved in the oral/practical M2 examinations. In order to create uniform and comparable testing practices among all of the teaching hospitals, at least one member of the university teaching staff co-administers the examinations at the teaching hospitals.

The scores assigned during the examinations should allow conclusions to be drawn about the competencies of the exam candidates. For this to be the case, grading must be based on fairness and equal opportunity, while also having the ability to stand up in court of law [4], [5], [6]. To ensure this, it is critical that a test's strengths and weaknesses are handled with awareness, consistent testing conditions and practices are ensured, content and expectations are competently structured and standardized, and the most objective evaluation criteria possible are applied [4], [7], [8], [9]. Doing justice to the requirements for quality criteria – objectivity, reliability and validity – is a known challenge facing oral and practical examinations [6], [10], [11], [12].

An effective measure to increase the quality, and with it the quality of practical and oral exams, are training programs for examiners [4], [8], [12], [13]. During training, strategies are imparted that can optimize not only the administration of tests, but also the evaluation of test performance [6].

In Baden-Württemberg a workshop for M2 examiners was designed in 2007 by the Competency Network for Teaching in Medicine (*Kompetenznetz "Lehre in der Medizin"*) [14] and is regularly held at all the medical schools in Baden-Württemberg to prepare examiners for the oral/practical part of the exam. This workshop consists of eight instructional units divided into seminar sessions and practical exercises and pursues the objectives of creating smooth testing procedures, optimizing tests based on experience, formulating test questions and tasks, and establishing criteria-based grading [7].

In a study by Öchsner, Geiler & Huber-Lang, examiners trained in the M2 workshops conducted at the Medical Faculty of Ulm were retrospectively surveyed about the effects and sustainability of the workshop using self-evaluation questionnaires [7]. The examiners were asked about the core issues of responding to strengths and weaknesses in the M2 exam with awareness, knowledge regarding the influence of reliability and validity on the oral/practical exam, confidence in designing test questions and following formal rules and regulations, and implementing a structured oral examination. The responses to all of the core issues pointed to benefits resulting from the examiner workshop, and these could be detected in the examiners even two years after completing the workshop. In close relation to the examiner perspectives presented in this study, our analysis here focuses on the output – the assignment of grades – at the Medical Faculty of Freiburg.

Aim

The aim of this study is to analyze the grading of the oral/practical M2 examination at the medical school in Freiburg in order to assess the internal quality assurance measures. The following four aspects were scrutinized:

- The correlation between written and oral/practical M2 grades;
- The difference in grading at the hospitals where testing was conducted: University Medical Center (UMC) versus the teaching hospitals;
- The difference in scoring between the examiners with and without prior training in one of the workshops;
- In-depth analysis of workshop sustainability in regard to grading.

Method

To analyze the assignment of grades for the oral/practical M2 exams, the Examination Office of Baden-Württemberg (*Landesprüfungsamt)* provided the anonymized data for the examinees in the 2012 spring and fall cohorts. Information was available on the candidates' written and oral/practical scores, as well as on the corresponding examination committees. In addition, it was known which examiners had previously completed the training workshop.

The size of the sample of examinees was N=315. The data set encompassed a total of 94 examination committees. An exam candidate was assigned to trained examiners for the analysis if one committee member had undergone training. The data were then entered and processed in SPSS (2012, version 20). Analysis was performed on the basis of the conventions of Bühner & Ziegler [15] and Bortz [16] and included descriptive statistical calculations and theory-based testing of hypotheses using inferential statistics.

Results

First, the correlation between the M2 examinees' scores on the written part and the oral/practical part was analyzed. The written scores displayed a mean value (M) of 2.45 (standard deviation (SD)=0.744). In contrast, a higher mean was calculated for the scores for the oral/practical part (M=1.92, SD=0.716). The grade of "very good" was assigned much more frequently for performance on the oral/practical part than for written test performance. Between the written and oral/practical scores, a slight linear correlation was seen with a highly significant rank correlation coefficient of ρ=0.460**.

It was then investigated if differences existed between the assignment of grades at the UMC and the other teaching hospitals. The mean and standard deviation for the UMC were M=1.95 and SD=0.768; for the teaching hospital these were M=1.88 and SD=0.661. Applying the Mann-Whitney U test, no significant difference between the two hospital groups was detected (ρ=0.682, Φ=0.023).

In addition, we checked for differences in the grading depending on if the oral/practical M2 exam was administered by examiners who had previously attended the workshop. The dichotomous variable was defined in that we evaluated whether or not an individual candidate was examined by a committee composed of examiners who

Table 1: Distribution by percentage

	Very good	Good	Satisfactory	Sufficient	Deficient
Written vs. Oral grades					
Written	6 %	50.6 %	36.7 %	5.4 %	1.3 %
Oral	26.7 %	56.8 %	15.2 %	0.3 %	1 %
UMC vs. Teaching hospitals					
UMC	24.6 %	61 %	11.9 %	0 %	2.5 %
Teaching hospitals	28.5 %	55.4 %	15.5 %	0.5 %	0 %
With vs. Without workshop					
With workshop	23.2 %	58.6 %	16.3 %	0.5 %	1.5 %
Without workshop	39.1 %	51.6 %	9.4 %	0	0
With vs. Without workshop, taking the time factor into account					
With workshop	17.9 %	61.6 %	17 %	0.9 %	2.7 %
Without workshop	38.3 %	50.4 %	11.3 %	0 %	0 %

Table 2: Descriptive statistics

Correlation hypothesis									
	N		M	SD	Min	Max	Spearman's rho		
Written vs. Oral grades									
Written	316		2.45	0.744	1	5	ρ = 0.460**		
Oral	315		1.92	0.716	1	5			
Difference hypothesis									
	N	N in %	M	SD	Min	Max	p	Z	Phi
UMC vs. Teaching hospitals									
UMC	118	37.5 %	1.95	0.768	1	5	0.682	-0.409	0.023
Teaching hospitals	193	61.3 %	1.88	0.661	1	4			
Missing values	4	1.2 %							
With vs. Without workshop									
With workshop	203	64.4 %	1.99	0.741	1	5	0.007	-2.708	0.165
Without workshop	64	20.9 %	1.70	0.634	1	3			
Missing values	48	14.7 %							
With vs. Without workshop, taking the time factor into account									
With workshop	116	36.7 %	2.08	0.782	1	5	0.000	-3.523	0.233
Without workshop	113	35.6 %	1.72	0.642	1	3			
Missing values	88	27.7 %							

had or had not undergone specific training. The analysis of the data for those with workshop training showed a mean of M=1.99 (SD=0.741); for those without the workshop the value is M=1.70 (SD=0.634). Using the Mann-Whitney U test, it could be shown that a highly significant difference of ρ=0.007**, with a weak effect of F=0.165, exists between the two groups regarding the distribution of the grades.

This calculation was carried out again in regard to the sustainability of the M2 examiner workshop. Here, the variable regarding the length of time between administering the exam and attending the workshop was evaluated. Examination committees were considered trained if workshop attendance had taken place no less than one year ago. For this group, the trained examiners displayed a mean of M=2.08 (SD=0.782). For the group of examiners who had not received training (or had attended the workshop over a year ago), the following value was determined: M=1.72 (SD=0.642). With the Mann-Whitney U test, a highly significant difference of ρ=0.000** was calculated with a small effect Φ=0.233.
A summary of these results is presented in Table 1and Table 2.

Discussion and Conclusions

The analysis of the grading was able to show that the efforts put into quality assurance at the Medical Faculty of Freiburg are worthwhile.
A highly significant correlation of ρ=0.460** was determined between the written and oral/practical grades. When focusing on the descriptive data analysis, it is seen that the scores assigned for oral/practical performance are on average better than those given for the written part. The grade of "very good" is assigned much more frequently for oral/practical performance than for written. This phenomenon has also been described in the literature [17].
The results of the other investigated aspects reveal no meaningful differences between the UMC and the other teaching hospitals concerning grading for the oral/practical M2 exam. This result can be viewed as an indication that the efforts to establish a uniform and comparable testing practice at all the hospitals are effective and should be retained at the Medical Faculty of Freiburg.
The results regarding the difference between candidates who were examined by committees with trained members or those without show a highly significant difference of p=0.007** with a small effect Φ=0.165. When looking carefully at the descriptive data, it is seen that committees with members who have attended the workshop give the grade of "very good" more seldom and assign a wider range of grades overall.
When examining the sustainability of the workshop (committees were defined as trained if the workshop had taken place less than one year prior), the result is even clearer: p=0.000** with the effect of Φ=0.233. The results presented here contrast with the study done by Öchsner, Geiler & Huber-Lang, in which evidence was found showing workshop sustainability over a period of two years on the basis of self-evaluation by the examiners [7]. Pertinent effects should be investigated in more detail in subsequent studies. Moreover, the role of the examination committee chairman should be included in the analysis. In the present study it was not possible to examine this aspect more closely.
Whether or not the observed change in examiners' grading after completing the workshop is accompanied by a higher quality of testing cannot be conclusively established. A positive effect of the workshop on grading does appear to be present, since the content imparted during the workshops is meant to assist in adequately assigning grades in terms of quality criteria.
In conclusion, it can be asserted that the M2 workshop helps achieve the highest quality possible for the oral/practical M2 examinations administered at the Medical Faculty of Freiburg. As a result, the M2 workshop can be recommended for all examiners [4], [7], [8], [9].
However, it must be pointed out that subject-specific testing cultures demonstrate a high level of stability over time when it comes to determining grades [4]. For this reason, it is recommended that any existing quality assurance measures regarding test quality be retained and intensified in order to effect long-term change in the testing practices related to grading. An indication for this could also be the higher significance for the analysis of the sustainability and the slight increase in effect size. In respect to this, a regular review of the knowledge gained in the M2 workshop should be considered; however, further studies are needed on the sustainability of acquired grading practices and of testing cultures in general.

Competing interests

The authors declare that they have no competing interests.

References

1. Bundesministerium für Gesundheit (BMG). Approbationsordnung für Ärzte: AÄppO, Teil I. Berlin: Bundesministerium für Gesundheit; 2002.
2. Krautter M, Jünger J, Koehl-Hackert N, Nagelmann L, Nikendei C. Evaluation eines strukturierten Prüfungsvorbereitungsprogramms für das 2. Staatsexamen (M2) nach Neuer Ärztlicher Approbationsordnung: Eine quantitative Analyse. ZFEQ. 2012;106(2):110–115. DOI: 10.1016/j.zefq.2011.09.020
3. Nikendei C, Weyrich P, Jünger J, Schrauth M. Medical Education in Germany. Med Teach. 2009;31(7):591–600. DOI: 10.1080/01421590902833010
4. Müller-Benedict V, Tsarouha E. Können Examensnoten verglichen werden? Eine Analyse von Einflüssen des sozialen Kontextes auf Hochschulprüfungen. Z Soziol. 2011;40(5):388–409.
5. Weber WD. Internationale Vergleichbarkeit von Noten im Hochschulbereich?: Problematik der Notenvergabe, Referenzgrößen und der Verwendung der Gauß'schen Normalverteilung. Qual Wissenschaft. 2010;4(1):20–23.
6. Fabry G. Medizindidaktik: Ein Handbuch für die Praxis. 1 Aufl. Programmbereich Medizin. Bern: Verlag Hans Huber; 2008.
7. Öchsner W, Geiler S, Huber-Lang M. Effekte und Nachhaltigkeit von Trainingsworkshops für den mündlich-praktischen Teil des M2-Examens. GMS Z Med Ausbild. 2013;30(3):Doc36. DOI: 10.3205/zma000879

8. Fischer MR, Holzer M, Jünger J. Prüfungen an den medizinischen Fakultäten - Qualität, Verantwortung und Perspektiven. GMS Z Med Ausbild. 2010;27(5):Doc66. DOI: 10.3205/zma000703

9. Schuwirth LW, van der Vleuten CP. Changing education, changing assessment, changing research? Med Educ. 2004;38(8):805–812. DOI: 10.1111/j.1365-2929.2004.01851.x

10. Khera N, Davies H, Davies H, Lissauer T, Skuse D, Wakeford, R, Stroobant J. How should paediatric examiners be trained? Arch Dis Child. 2005;90(1):43–47. DOI: 10.1136/adc.2004.055103

11. Seyfarth M, Reincke M, Seyfarth J, Ring J, Fischer MR. Neue ärztliche Approbationsordnung und Notengebung beim Zweiten Staatsexamen: Eine Untersuchung an zwei bayerischen medizinischen Fakultäten. Dtsch Arztbl Int. 2010;28-29:500–504. DOI: 10.3238/arztebl.2010.0500

12. Wakeford R, Southgate L, Wass V. Improving oral examinations: selecting, training, and monitoring examiners for the MRCGP: Royal College of General Practitioners. BMJ. 1995;311:931–935. DOI: 10.1136/bmj.311.7010.931

13. Schuwirth L. The need for national licensing examinations. Med Educ. 2007;41(11):1022–1023. DOI: 10.1111/j.1365-2923.2007.02856.x

14. Fegert J, Obertacke U, Resch F, Hilzenbecher M. Die Qualität der Lehre nicht dem Zufall überlassen. Dtsch Arztebl. 2009;106(7):290–291.

15. Bühner M, Ziegler M. Statistik für Psychologen und Sozialwissenschaftler. Always learning. München: Pearson Studium; 2012.

16. Bortz J. Statistik für Human- und Sozialwissenschaftler. 6 Aufl. Heidelberg: Springer-Verlag; 2005.

17. Bussche van HD, Wegscheider K, Zimmermann T. Medizinische Fakultäten: Der Ausbildungserfolg im Vergleich (III). Dtsch Arztbl Int. 2006;103(47):B-2762/C-2644.

Corresponding author:
Peter Brüstle
Uni Freiburg, Kompetenzzentrum Lehrevaluation in der Medizin Baden-Württemberg, Sitz Freiburg, Freiburg, Deutschland
peter.bruestle@uniklinik-freiburg.de

Please cite as
Schickler A, Brüstle P, Biller S. The Final Oral/Practical State Examination at Freiburg Medical Faculty in 2012 – Analysis of grading to test quality assurance. GMS Z Med Ausbild. 2015;32(4):Doc39. DOI: 10.3205/zma000981, URN: urn:nbn:de:0183-zma0009816

This article is freely available from
http://www.egms.de/en/journals/zma/2015-32/zma000981.shtml

Received: *2014-01-14*
Revised: *2014-10-23*
Accepted: *2015-01-05*
Published: *2015-10-15*

OPEN ACCESS

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

Mündlich-praktischer Teil des Zweiten Abschnitts der Ärztlichen Prüfung in Freiburg 2012 – Analyse der Notenvergabe zur Überprüfung von Qualitätssicherungsmaßnahmen

Angela Schickler[1]
Peter Brüstle[1]
Silke Biller[1]

1 Uni Freiburg, Kompetenzzentrum Lehrevaluation in der Medizin Baden-Württemberg, Sitz Freiburg, Freiburg, Deutschland

Zusammenfassung

Zielsetzung: Von der Medizinischen Fakultät Freiburg werden verschiedene Maßnahmen zur Qualitätssicherung des mündlich-praktischen Teils des Zweiten Abschnitts der Ärztlichen Prüfung (M2) betrieben. Insbesondere werden die Prüfenden in einem Baden-Württemberg-weit standardisierten M2-Prüferworkshop fortgebildet.
Ziel der vorliegenden Studie war es die Notenvergabe zu analysieren. Es wurde geprüft, welcher Zusammenhang zwischen der schriftlichen und der mündlich-praktischen Notenvergabe besteht und ob sich ein Unterschied in der Notenvergabe zwischen Universitätsklinikum (UKL) und den Akademischen Lehrkrankenhäusern (ALKs) findet. Darüber hinaus wurde untersucht, ob sich die Notenvergabe von fortgebildeten und nicht fortgebildeten Prüfenden unterscheidet und inwieweit diese Unterschiede sich im zeitlichen Abstand zum Workshop verändern.
Methodik: Die analysierte Stichprobe (N=315) umfasst die Frühjahrs- und Herbstprüfungskohorte 2012. Der Datensatz wurde vom Landesprüfungsamt zur Verfügung gestellt und mittels deskriptiver und Inferenzstatistik ausgewertet.
Ergebnisse: Zwischen der schriftlichen und mündlich-praktischen M2-Prüfungsnote konnte ein Zusammenhang von ρ=0,460** ermittelt werden. Es konnte kein signifikanter Unterschied in der Verteilung der Noten zwischen UKL und den ALKs festgestellt werden. Die Daten zu den Variablen fortgebildete versus nicht fortgebildete Prüfende zeigen, dass Prüfende mit Absolvierung eines Prüferworkshops seltener die Note „sehr gut“ sowie ein breiteres Notenspektrum vergeben. In der Verteilung der Noten zeigt sich ein signifikanter Unterschied (p=0,007, phi=0,165). Dieses Ergebnis wird deutlicher bei Betrachtung von Prüfenden, deren Workshopteilnahme nicht länger als ein Jahr zurücklag.
Schlussfolgerungen: Die Ergebnisse der Analyse weisen darauf hin, dass der Prüferworkshop an der Medizinischen Fakultät Freiburg eine wirksame Qualitätssicherungsmaßnahme ist. Die Ergebnisse lassen den Schluss zu, die Prüfenden zur Absolvierung eines Prüferworkshops zu motivieren.

Schlüsselwörter: Staatsexamen Medizin, mündlich-praktische M2-Prüfung, Prüferworkshop, Notengebung

Hintergrund

In der Ärztlichen Approbationsordnung von 2002 [1] kam es zu grundlegenden Änderungen des Staatsexamens. So bestand dieses bis zu den aktuellen Änderungen von 2012 aus zwei Abschnitten. Der erste Abschnitt (M1) wurde nach den ersten beiden Studienjahren abgelegt, der zweite Abschnitt (M2) nach dem gesamten Studium inklusive des Praktischen Jahres (PJ). Beide Prüfungsabschnitte unterteilten sich in einen schriftlichen und einen mündlich-praktischen Teil. Einige der Änderungen von 2002 betrafen den mündlich-praktischen Teil des M2: Die mündlich-praktische M2-Prüfungsnote wiegt in der Abschlussgesamtnote der Studierenden ein Drittel. Die Prüfungszeit wurde auf zwei ganze Tage verlängert und ein praktischer Prüfungsanteil wurde explizit eingeführt. Insgesamt führen vier bis fünf Prüfende die Prüfung von maximal vier Prüflingen durch [1]. Es wird gefordert, den Schwerpunkt dieser Prüfung auf patientenbezogene Fragestellungen zu legen [1], [2], [3].

Grundsätzlich handelt es sich bei der mündlich-praktischen M2-Prüfung um eine sehr aufwendige berufsqualifizierende Prüfung, welche für die Prüfenden mit erheblichen Anforderungen verbunden ist. Der Medizinische Fakultätentag (MFT) hat 2011 in einer Resolution auf die hohe Belastung der Medizinischen Fakultäten hingewiesen. Er gab zu bedenken, dass durch diesen erhöhten Ressourceneinsatz und die dadurch notwendige Ausweitung des Pools an Prüfenden auch auf Personal außerhalb der Universitäten ein Qualitätsverlust der Prüfungen und eine Ungleichbehandlung der Prüflinge möglich seien [http://www.mft-online.de/files/200_omft_2011.pdf zuletzt geprüft am 05.11.2013]. Die Medizinische Fakultät Freiburg arbeitet bei der Ausbildung der PJ-Studierenden mit 15 Akademischen Lehrkrankenhäusern (ALKs) zusammen, welche auch an den mündlich-praktischen M2-Prüfungen beteiligt sind. Um zwischen allen Ausbildungskrankenhäusern eine einheitliche und gleichwertige Prüfungspraxis herzustellen, nimmt mindestens eine Person des universitären Lehrkörpers die Prüfungen an den ALKs mit ab.

Die bei der Prüfung vergebenen Noten sollen Rückschlüsse auf die Kompetenzen der Prüfungskandidierenden zulassen. Dazu muss die Notenvergabe auf Fairness und Chancengleichheit basieren, aber auch Gerichtsfestigkeit aufweisen [4], [5], [6]. Um dies zu gewährleisten, sind ein bewusster Umgang mit den Stärken und Schwächen der Prüfung, Konstanz in den Prüfungsbedingungen und der Prüfungspraxis, eine gute Strukturierung und Standardisierung der Inhalte sowie Erwartungshorizonte und möglichst objektive Bewertungskriterien für die Prüfung notwendig [4], [7], [8], [9]. Den Anforderungen an die Gütekriterien – Objektivität, Reliabilität und Validität – gerecht zu werden ist bei mündlich-praktischen Prüfungen ein bekanntes Problem [6], [10], [11], [12].

Eine wirksame Maßnahme um die Güte und damit einhergehend die Qualität von mündlich-praktischen Prüfungen zu steigern sind Fort- und Weiterbildungsmaßnahmen der Prüfenden [4], [8], [12], [13]. Inhaltlich werden dabei Maßnahmen und Strategien vermittelt, die sowohl die Vorbereitung als auch die Durchführung von Prüfungen sowie auch die Bewertung der Prüfungsleistungen optimieren können [6].

In Baden-Württemberg wurde 2007 vom Kompetenznetz „Lehre in der Medizin“ [14] ein M2-Prüferworkshop konzipiert, der an allen baden-württembergischen Fakultäten zur Vorbereitung der Prüfenden auf den mündlich-praktischen Teil regelmäßig durchgeführt wird. Der acht Unterrichtseinheiten umfassende Workshop untergliedert sich in Seminaranteile und praktische Übungen. Hier werden die Ziele „reibungsloser Prüfungsablauf“, „Prüfungen auf Grund bestehender Erfahrungen optimieren“, „Prüfungsfragen und –aufgaben formulieren“ sowie „kriterienbezogene Notenvergabe etablieren“ verfolgt [7].

In einer Studie von Öchsner, Geiler und Huber-Lang wurden mittels Selbsteinschätzungsfragebogen die durch M2-Prüferworkshops fortgebildeten Prüfenden an der Medizinischen Fakultät in Ulm retrospektiv zu den Effekten und der Nachhaltigkeit des Prüferworkshops befragt [7]. Die Prüfenden wurden zu den Kernfragestellungen „bewusster Umgang mit Stärken und Schwächen der M2-Prüfung“, „Kenntnisse über den Einfluss der Reliabilität und Validität auf die mündlich-praktische-Prüfung“, „Sicherheit bei der Aufgabenkonstruktion und Prüfungsregularien“ und „Umsetzung des Konzepts des strukturierten mündlichen Prüfens“ befragt. Die Ergebnisse zu allen Kernfragestellungen lassen auf einen Nutzen des Prüferworkshops schließen und konnten bei den Prüfenden auch zwei Jahre nach der Absolvierung des Prüferworkshops nachgewiesen werden. Angrenzend an die in dieser Studie beleuchtete Perspektive der Prüfenden soll bei der vorliegenden Analyse der Fokus auf den Output, die Notengebung an der Medizinischen Fakultät in Freiburg, gelegt werden.

Zielsetzung

Ziel der vorliegenden Studie war es, die Notenvergabe zur mündlich-praktischen M2-Prüfung an der Medizinischen Fakultät Freiburg zu analysieren, um die internen Qualitätssicherungsmaßnahmen zu prüfen. Hierfür wurde folgenden vier Fragestellungen nachgegangen:

- Zusammenhang zwischen schriftlichen und mündlich-praktischen M2-Prüfungsnoten
- Unterschied in der Notengebung an den Prüfungskrankenhäusern: UKL versus ALKs
- Unterschied in der Notengebung zwischen den Prüfenden mit und ohne voriger Absolvierung eines Prüferworkshop
- Vertiefte Analyse zur Nachhaltigkeit des Prüferworkshop bezüglich der Notengebung.

Methodik

Zur Analyse der Notengebung der mündlich-praktischen M2- Prüfung wurden die Daten der Prüfungskandidierenden der Frühjahrs- und Herbstkohorte 2012 vom Landesprüfungsamt für Medizin und Pharmazie Baden-Württemberg in anonymisierter Form zur Verfügung gestellt. Es lagen Angaben zu den schriftlichen und mündlich-praktischen Noten der Prüflinge sowie zu den entsprechenden Prüfungskommissionen der mündlich-praktischen Prüfung vor. Zudem war bekannt, welche Prüfenden zuvor einen Prüferworkshop absolviert hatten.

Die Stichprobe der Prüflinge umfasste eine Fallzahl von N=315. Der Datensatz umfasste insgesamt 94 Prüfungskommissionen. Ein Prüfling wurde bei der Analyse geschulten Prüfenden zugeordnet, sobald ein Kommissionsmitglied geschult war. Die Daten wurden in SPSS (2012, Version 20) eingegeben und dort verarbeitet. Die Auswertung erfolgte auf der Grundlage der Konventionen von Bühner und Ziegler [15] und Bortz [16] und umfasste deskriptive statistische Berechnungen und theoriegeleitete Hypothesenprüfung mittels Inferenzstatistik.

Ergebnisse

Zuerst wurde der Zusammenhang zwischen den schriftlichen Noten und den mündlich-praktischen Noten der M2-Prüflinge untersucht. Bei den schriftlichen Noten lag der Mittelwert (M) bei 2,45 (Standardabweichung (SD)=0,744). Bei den mündlich-praktischen Noten hingegen, wurde ein besserer Mittelwert (M=1,92, SD=0,716) ermittelt. Die Note „sehr gut“ wurde erheblich häufiger mündlich-praktisch vergeben als schriftlich. Zwischen der schriftlichen und der mündlich-praktischen Note zeigte sich ein schwacher linearer Zusammenhang mit einem hochsignifikanten Rangkorrelationskoeffizient von ρ=0,460**.

Anschließend wurde untersucht, ob sich Unterschiede in der Notenvergabe zwischen dem UKL und den ALKs finden.

Der Mittelwert und die Standardabweichung vom UKL lagen bei M=1,95 und SD=0,768, bei den ALKs bei M=1,88 und SD=0,661. Mittels Mann-Whitney-U-Test wurde kein signifikanter Unterschied zwischen den beiden Gruppen der Prüfungskrankenhäuser ermittelt (ρ=0,682, Φ=0,023).

Ferner wurde der Fragestellung nachgegangen, ob sich Unterschiede in der Benotung finden, wenn die mündlich-praktische M2-Prüfung von Prüfenden abgenommen wurde, die zuvor einen Prüferworkshop absolviert haben. Die dichotome Variable wurde mittels Betrachtung des einzelnen Prüflings gebildet, indem bewertet wurde, ob diese/-r durch eine Prüfungskommission mit oder ohne fortgebildeten Prüfenden geprüft wurde. Die Analyse der Daten für die Ausprägung mit Prüferworkshop zeigt einen Mittelwert von M=1,99 (SD=0,741), ohne Prüferworkshop ist der Wert M=1,70 (SD=0,634). Mittels Mann-Whitney-U-Test konnte gezeigt werden, dass zwischen den beiden Gruppen ein hochsignifikanter Unterschied in der Verteilung der Noten von ρ=0,007** besteht – bei einem schwachen Effekt von Φ=0,165.

Erneut wurde diese Rechnung in Bezug auf die Nachhaltigkeit des M2-Prüferworkshops durchgeführt. Hierzu wurde die Variable bezüglich des zeitlichen Abstandes zwischen Prüfung und Prüferworkshops bewertet. Dabei galten Kommissionen als geschult, wenn der Workshop nicht länger als maximal ein Jahr zurück lag. Bei dieser Gruppe der geschulten Prüfenden zeigte sich ein Mittelwert von M=2,08 (SD=0,782). Bei der Gruppe der Prüfenden, die nicht geschult waren (bzw. bei denen der Workshop länger als ein Jahr zurück lag) ergab sich folgender Wert: M=1,72 (SD=0,642). Mittels Mann-Whitney-U-Test konnte ein hochsignifikanter Unterschied von ρ=0,000** errechnet werden mit einem kleinen Effekt von Φ=0,233. Die zusammenfassende Darstellung der Ergebnisse findet sich in Tabelle 1 und Tabelle 2.

Diskussion und Schlussfolgerungen

Die Analyse der Notenvergabe konnte zeigen, dass sich die Bemühungen der Qualitätssicherung an der Medizinischen Fakultät Freiburg lohnen.

Es konnte ein hochsignifikanter Zusammenhang von ρ=0,460** zwischen schriftlicher und mündlich-praktischer Notengebung ermittelt werden. Bei Betrachtung der deskriptiven Datenanalyse zeigt sich, dass die mündlich-praktisch vergebenen Noten im Vergleich zu den schriftlichen im Mittel besser ausfallen. Die Note „sehr gut“ wird erheblich häufiger mündlich-praktisch als schriftlich vergeben. Dies entspricht einem auch in der Literatur beschriebenem Phänomen [17].

Die Ergebnisse der weiteren Fragestellung zeigen keinen bedeutsamen Unterschied in der Benotung der mündlich-praktischen M2-Prüfung zwischen dem UKL und den ALKs. Dieses Ergebnis kann als Hinweis gewertet werden, dass sich die unternommenen Anstrengungen, zwischen allen Prüfungskrankenhäusern eine einheitliche und gleichwertige Prüfungspraxis herzustellen, lohnen und somit folglich an der Medizinischen Fakultät Freiburg weiter beibehalten werden sollten.

Die Ergebnisse bezüglich des Unterschieds zwischen Prüflingen, die von Prüfungskommissionen mit oder ohne fortgebildeten Mitgliedern geprüft wurden zeigen einen hochsignifikanten Unterschied von ρ=0,007** mit einem schwachen Effekt Φ=0,165. Bei der Betrachtung der deskriptiven Daten zeigt sich, dass Kommissionen mit Mitgliedern die einen Prüferworkshop absolvierten seltener die Note „sehr gut“ sowie ein generell breiteres Notenspektrum vergeben.

Bei Betrachtung der Nachhaltigkeit des Workshops (Kommissionen galten als geschult, wenn der Workshop nicht länger als maximal ein Jahr zurück lag) wird das Ergebnis noch deutlicher: ρ=0,000** mit dem Effekt von Φ=0,233. Die hier dargestellten Ergebnisse stehen in Kontrast zur Studie von Öchsner, Geiler und Huber-Lang. Dort wurden über einen Zeitraum von zwei Jahren anhand der Selbsteinschätzung der Prüfenden Hinweise für die Nachhaltigkeit des Prüferworkshops gewonnen [7]. Entsprechende Effekte gilt es in Folgestudien genauer zu analysieren. Hierbei sollte zusätzlich auch die Rolle der Prüfungsvorsitzenden in die Analyse aufgenommen werden. In dieser Studie konnte dieser Aspekt nicht näher untersucht werden.

Ob die festgestellte veränderte Notenvergabe nach Absolvierung eines Prüferworkshops mit einer höheren Güte der Prüfung einhergeht, lässt sich nicht abschließend klären. Ein positiver Effekt des Prüferworkshops auf die Notenvergabe erscheint begründet, da die dort gelehrten Inhalte den Prüfenden zu adäquater Notenfindung im Sinne der Gütekriterien helfen sollen.

Schlussfolgernd kann festgehalten werden, dass das Ziel, an der Medizinischen Fakultät Freiburg eine höchstmögliche Qualität der mündlich-praktischen M2-Prüfung zu erreichen, durch die M2-Prüferworkshops unterstützt werden kann. Somit lässt sich der M2-Prüferworkshop für alle Prüfenden empfehlen [4], [7], [8], [9].

Tabelle 1: Prozentuale Verteilung

	Sehr gut	gut	befriedigend	ausreichend	mangelhaft
Schriftlich vs. mündliche Note					
schriftlich	6 %	50,6 %	36,7 %	5,4 %	1,3 %
mündlich	26,7 %	56,8 %	15,2 %	0,3 %	1 %
UKL vs. ALKs					
UKL	24,6 %	61 %	11,9 %	0 %	2,5 %
ALKs	28,5 %	55,4 %	15,5 %	0,5 %	0 %
Mit vs. ohne Prüferworkshop					
Mit Prüferworkshop	23,2 %	58,6 %	16,3 %	0,5 %	1,5 %
Ohne Prüferworkshop	39,1 %	51,6 %	9,4 %	0	0
mit vs. ohne Prüferworkshop unter Berücksichtigung des zeitlichen Faktors					
Mit Prüferworkshop	17,9 %	61,6 %	17 %	0,9 %	2,7 %
Ohne Prüferworkshop	38,3 %	50,4 %	11,3 %	0 %	0 %

Tabelle 2: Deskriptive Statistik

Zusammenhangshypothese									
	N		M	SD	Min	Max	Spearman Rho		
Schriftlich vs. mündliche Note									
schriftlich	316		2,45	0,744	1	5	ρ = 0,460**		
mündlich	315		1,92	0,716	1	5			
Unterschiedshypothese									
	N	N in %	M	SD	Min	Max	p	Z	Phi
UKL vs. ALKs									
UKL	118	37,5 %	1,95	0,768	1	5	0,682	-0,409	0,023
ALKs	193	61,3 %	1,88	0,661	1	4			
Fehlende Werte	4	1,2 %							
Mit vs. ohne Prüferworkshop									
Mit Prüferworkshop	203	64,4 %	1,99	0,741	1	5	0,007	-2,708	0,165
Ohne Prüferworkshop	64	20,9 %	1,70	0,634	1	3			
Fehlende Werte	48	14,7%							
mit vs. ohne Prüferworkshop unter Berücksichtigung des zeitlichen Faktors									
mit Prüferworkshop	116	36,7 %	2,08	0,782	1	5	0,000	-3,523	0,233
ohne Prüferworkshop	113	35,6 %	1,72	0,642	1	3			
fehlende Werte	88	27,7 %							

Es muss jedoch darauf hingewiesen werden, dass fachspezifische Prüfungskulturen zur Notengebung eine hohe zeitliche Stabilität aufweisen [4]. Aus diesem Grund ist zu empfehlen, eingeführte Qualitätssicherungsmaßnahmen bezüglich der Prüfungsqualität beizubehalten und zu intensivieren, um eine langfristige Änderung der Prüfungspraktiken im Hinblick auf die Notenfindung zu bewirken. Auch die Intensivierung der Signifikanz bei der Untersuchung der Nachhaltigkeit sowie der leichte Anstieg der Effektstärke können ein Indiz hierfür sein. So sollte man diesbezüglich über eine regelmäßige Auffrischung der Kenntnisse aus dem M2-Prüferworkshop nachdenken. Allerdings sind weitere Untersuchungen zur Nachhaltigkeit

von erlernten Praktiken bei der Notengebung und von allgemeinen Prüfungskulturen notwendig.

Interessenkonflikt

Die Autoren erklären, dass sie keine Interessenkonflikte im Zusammenhang mit diesem Artikel haben.

Literatur

1. Bundesministerium für Gesundheit (BMG). Approbationsordnung für Ärzte: AÄppO, Teil I. Berlin: Bundesministerium für Gesundheit; 2002.
2. Krautter M, Jünger J, Koehl-Hackert N, Nagelmann L, Nikendei C. Evaluation eines strukturierten Prüfungsvorbereitungsprogramms für das 2. Staatsexamen (M2) nach Neuer Ärztlicher Approbationsordnung: Eine quantitative Analyse. ZFEQ. 2012;106(2):110–115. DOI: 10.1016/j.zefq.2011.09.020
3. Nikendei C, Weyrich P, Jünger J, Schrauth M. Medical Education in Germany. Med Teach. 2009;31(7):591–600. DOI: 10.1080/01421590902833010
4. Müller-Benedict V, Tsarouha E. Können Examensnoten verglichen werden? Eine Analyse von Einflüssen des sozialen Kontextes auf Hochschulprüfungen. Z Soziol. 2011;40(5):388–409.
5. Weber WD. Internationale Vergleichbarkeit von Noten im Hochschulbereich?: Problematik der Notenvergabe, Referenzgrößen und der Verwendung der Gauß'schen Normalverteilung. Qual Wissenschaft. 2010;4(1):20–23.
6. Fabry G. Medizindidaktik: Ein Handbuch für die Praxis. 1 Aufl. Programmbereich Medizin. Bern: Verlag Hans Huber; 2008.
7. Öchsner W, Geiler S, Huber-Lang M. Effekte und Nachhaltigkeit von Trainingsworkshops für den mündlich-praktischen Teil des M2-Examens. GMS Z Med Ausbild. 2013;30(3):Doc36. DOI: 10.3205/zma000879
8. Fischer MR, Holzer M, Jünger J. Prüfungen an den medizinischen Fakultäten - Qualität, Verantwortung und Perspektiven. GMS Z Med Ausbild. 2010;27(5):Doc66. DOI: 10.3205/zma000703
9. Schuwirth LW, van der Vleuten CP. Changing education, changing assessment, changing research? Med Educ. 2004;38(8):805–812. DOI: 10.1111/j.1365-2929.2004.01851.x
10. Khera N, Davies H, Davies H, Lissauer T, Skuse D, Wakeford, R, Stroobant J. How should paediatric examiners be trained? Arch Dis Child. 2005;90(1):43–47. DOI: 10.1136/adc.2004.055103
11. Seyfarth M, Reincke M, Seyfarth J, Ring J, Fischer MR. Neue ärztliche Approbationsordnung und Notengebung beim Zweiten Staatsexamen: Eine Untersuchung an zwei bayerischen medizinischen Fakultäten. Dtsch Arztbl Int. 2010;28-29:500–504. DOI: 10.3238/arztebl.2010.0500
12. Wakeford R, Southgate L, Wass V. Improving oral examinations: selecting, training, and monitoring examiners for the MRCGP: Royal College of General Practitioners. BMJ. 1995;311:931–935. DOI: 10.1136/bmj.311.7010.931
13. Schuwirth L. The need for national licensing examinations. Med Educ. 2007;41(11):1022–1023. DOI: 10.1111/j.1365-2923.2007.02856.x
14. Fegert J, Obertacke U, Resch F, Hilzenbecher M. Die Qualität der Lehre nicht dem Zufall überlassen. Dtsch Arztebl. 2009;106(7):290–291.
15. Bühner M, Ziegler M. Statistik für Psychologen und Sozialwissenschaftler. Always learning. München: Pearson Studium; 2012.
16. Bortz J. Statistik für Human- und Sozialwissenschaftler. 6 Aufl. Heidelberg: Springer-Verlag; 2005.
17. Bussche van HD, Wegscheider K, Zimmermann T. Medizinische Fakultäten: Der Ausbildungserfolg im Vergleich (III). Dtsch Arztbl Int. 2006;103(47):B-2762/C-2644.

Korrespondenzadresse:
Peter Brüstle
Uni Freiburg, Kompetenzzentrum Lehrevaluation in der Medizin Baden-Württemberg, Sitz Freiburg, Freiburg, Deutschland
peter.bruestle@uniklinik-freiburg.de

Bitte zitieren als
Schickler A, Brüstle P, Biller S. The Final Oral/Practical State Examination at Freiburg Medical Faculty in 2012 – Analysis of grading to test quality assurance. GMS Z Med Ausbild. 2015;32(4):Doc39. DOI: 10.3205/zma000981, URN: urn:nbn:de:0183-zma0009816

Artikel online frei zugänglich unter
http://www.egms.de/en/journals/zma/2015-32/zma000981.shtml

Eingereicht: *14.01.2014*
Überarbeitet: *23.10.2014*
Angenommen: *05.01.2015*
Veröffentlicht: *15.10.2015*