Predictive validity of a tool to resolve borderline grades in OSCEs

Abstract

There is inconclusive evidence suggesting which standard setting method yields the highest validity for pass/fail decisions in examinations. The Objective Borderline Method 2 (OBM2) is a decision-making tool for reclassification of borderline grades to clear pass or clear fail grades to resolve examiner uncertainty for high-stakes pass/fail decisions.

This study evaluated the predictive validity of OBM2 pass/fail decisions, using consecutive years' Objective Structured Clinical Examination (OSCE) results within a medical cohort (n=271) at the University of New South Wales, Australia. OBM2 decisions in one OSCE (n=687) were compared to marks obtained in a subsequent OSCE via independent samples T-tests and analysis of variance (ANOVA). The extent of the relationship between these two variables determines the predictive validity of OBM2 decisions, given that past student grades are capable of predicting future performance.

OBM2 decisions in an initial OSCE were found to have a statistically significant predictive nature for subsequent OSCE marks (p=.005). For initial decisions which reclassified to a pass grade, subsequent OSCE marks were significantly higher than for the cases where initial decisions were reclassified to a fail grade. Stronger associations were identified between related assessment domains/criteria compared to unrelated domains/criteria (Cohen's d=.469 vs Cohen's d=.388 respectively).

Through demonstrating the OBM2 decisions' predictive association across exams there is support for the OBM2's predictive validity, deeming it a promising method to be used for resolving examiner uncertainty when making pass/fail decisions within OSCEs.

Keywords: OCSE, borderline grades, assessment, medical students

Rowan Klein Nulend[1]
Peter Harris[1]
Boaz Shulruf[1]

1 University of New South Wales, Office of Medical Education, Sydney, Australia

1. Introduction

It is important that any decisions arising from assessment strategies used within a medical program are defensible [1], [2], [3]. Subjectivity reduces the defensibility of an examination; to increase objectivity in OSCE settings it is common that a standard setting method is applied [4]. Standard setting methods are applied to define cut-scores which correspond to a minimum level of proficiency/achievement required in an assessment task [4], [5].

A broad range of standard setting methods exists; all methods explored in current literature feature some subjectivities and imprecisions, with inconclusive evidence surrounding their efficacy [6], [7]. Most methods require judgements of experts/judges. Although these judgements are made by experts in the field, it is impossible to be entirely objective in such instances [6], [8], [9]. Since there is no gold-standard for standard setting, validating a standard setting is the most challenging issue in standard setting [8]. Previous studies demonstrated that when two or more standard setting methods are applied to the same data set, each delivers a different cut-score [10], [11], [12].

Another issue is the definition of a borderline or, as commonly described "minimally competent" student, and the variability of expert opinions in this domain [13]. A borderline result is observed when the examiner is uncertain whether the observed performance reached the clear pass or clear fail level. This may occur when student's observed performance lies near the expected cut-score which distinguishes between the pass and fail grades [4]. To resolve this issue the Objective Borderline Method (OBM) was introduced [10]. The OBM is a standard setting method which uses the concept of redefining borderline marks into either a pass or fail grade; derived from the proportions of pass, borderline and fails yielded by all examinees [14]. This model is based on probability, using proportions of pass/borderline/fail marks. Instead, most standard setting methods allocate a cut-score based on expert opinion or statistical techniques, as is done with the Angoff method and borderline regression method respectively [10].

Since the introduction of the OBM, the Objective Borderline Method 2 (OBM2) has been developed. The OBM2

is not a standard-setting method, as it does not establish a cut-score. The OBM2, instead, is a decision-making tool for reclassification of borderline grades. It uses only two measures; examinee ability and item difficulty, estimated from all assessment marks from an exam, to reclassify the borderline grade as either pass or fail on a case-by-case basis. The OBM2 was found applicable within standard clinical style examination settings to support pass or fail grade decisions in borderline instances [15]. The OBM2 is a probability based method used to replace a borderline mark with either pass of fail mark given to an examinee for each single item [16], [17]. Thus, an examinee may receive any number of borderline marks, from zero to the total number of items in the examination (in the current study it may span between 0 to 54 per student). A borderline mark is a mark given to the examinee when the examiner is unable to determine that a particular skill was performed either at the clear pass or clear fail level [16], [17]. The reclassification of the borderline marks to either pass or fail is determined by the proportions of passes (p), borderline (b) and fail (f) marks yielded by the students using the formula: "OBM index=(p/[b+p])×(b/[f+b])" [16]. The OBM index is calculated twice; once for marks of all items yielded by the student to determine "student ability", and once for all marks yielded by each item by all students to determine "item difficulty". Thus, for every borderline mark there are two OBM indices. Then the OBM indices are compared for a given borderline mark. If "student ability"≥ "item difficulty", the borderline mark is reclassified to a pass. If "student ability"<"item difficulty" the borderline mark is reclassified to a fail. A detailed explanation of the technicality of the OBM2 in presented in previous research [16].

In the setting of education, predictive validity is an important subset of criterion validity, as an important goal of examinations is to predict future performance [18]. Current literature indicates that past student grades predict future performance [19]. If the OBM2 could reflect this expectation within a group of students who have all been allocated the same mark (borderline) and had this reclassified to a pass or a fail, it would enhance the OBM2's validity as a tool to reclassify borderline grades to either "clear pass" or "clear fail" grades. That is, does the OBM2 decision place a borderline student into a group where their future performance corresponds with what is expected from students, based on past grades.

Previous studies have explained the OBM2 tool and have assessed the tool's defensibility, feasibility, impact on OSCE results and validity [14], [16], [17]. However, these studies used snapshot data which could not provide any indication of the predictive validity of the OBM2 pass/fail decisions [10], [14], [16].

2. Aim

The aim of this study was to determine to what extent decisions made by the OBM2 predict future performance. This may determine the predictive validity of pass/fail decisions made by the OBM2. To achieve this, the following research question was used: what is the extent of the association between OBM2 decisions in one OSCE with the marks obtained in a subsequent year's OSCE?

3. Study setting

This study uses data from OSCEs conducted at the University of New South Wales (UNSW) in Sydney, Australia. UNSW medicine is a six year undergraduate program and has OSCEs in second year, third year and sixth year [20]. This study uses data from year 2 OSCE (referred to as Initial) and year 3 OSCE (referred to as Subsequent) examinations of the same cohort, in two consecutive years (2016-2017). The first two years of the UNSW medicine program are primarily theoretical, with weekly alternating 2-hour long clinical skills sessions on-campus and in the hospital being students' sole clinical practice. Meanwhile, third year students are placed at an allocated hospital daily throughout the year, allowing students substantially more clinical training [17], [18].

The initial examination assesses students (n=271) across three domains; general communication, clinical communication and physical examination, which are split into nine specific assessment criteria within the marking rubric. Therefore, a student is able to achieve up to nine borderline results per OSCE station. The cohort is divided across four separate sites [21]. The subsequent examination (257 students) uses slightly different assessment criteria (see table 1) [21] and is conducted across nine separate sites.

Both the initial and subsequent OSCEs consist of six separate stations, with different cases and examiners [21]. Each station has one examiner, with a mix of external and university-affiliated examiners. The initial OSCE allows fifteen minutes per station and emphasises assessment of clinical skills, such as clinical communication, physical examination and general communication [21]. The subsequent OSCE allows ten minutes per station and relies on similar clinical skills, as well as case-specificity; meaning thorough underlying clinical knowledge is necessary to perform well in the examination [21]. These subsequent criteria each have equivalents to the three initial domains and can therefore be compared. Both the initial and subsequent OSCEs allow for one re-attempt after a fail grade. Examiners for the subsequent OSCEs were not aware of student grades yielded in the initial OSCE.

The study comprised data of 271 students who completed the year 2 OSCE in 2016. The year 2 OSCE consists of six stations, in each of which the student is assessed by nine assessment criteria, resulting with 54 marks per student in year 2 OSCE. Each of the assessment criteria focus on one of the three domains; general communication, clinical communication, or physical examination. In total year 2 OSCE yielded 14,634 marks (f=83 [0.6%]; b=687 [4.7%]; p=13864 [94.7%], the p mark includes both "pass" and "distinction" marks). After the application

Table 1: Subsequent Overall Assessment Criteria Generated by Clinical Experts

Assessment Criterion History Based Station	Assessment Criterion Physical Examination Based Station	Unified Criterion
Perform technically competent physical examination or skill	Detect physical signs	Physical Examination
Elicit a relevant clinical history	Gather a relevant psychosocial, past medical and family history	History
Listen attentively, engage patient and maintain respect	Engage patient and maintain respect	Communication
Interpret patient history and presentation	Interpret patient case	Case Interpretation
Summarise history to the examiner	Summarise case findings	Case Summary

of the OBM2, which replaced the borderline marks with either passes or fails, the marks are summarised (averaged) by the three domains and reported as such. This study however, focussed only on the 687 borderline marks, since only these were modified to either pass of fail.

4. Methods

Hereafter, "OBM2 decisions to reclassify borderline grades to either clear pass or clear fail grades" will be referred to as "decisions".

One data set included all initial borderline results for which decisions were made (n=687); the second included all the subsequent marks correlating to each initial decision. For 58 of the 687 initial borderline decisions (14 students), the subsequent OSCE was not attempted in the consecutive year, meaning these subsequent entries were incomplete and were excluded from the analysis. Therefore 629 sets of decisions (257 students) were analysed. During the initial examination, a student can receive a maximum of nine borderline results per OSCE station, as there are nine criteria according to which students are assessed within each station.

The subsequent data consisted of the original marks across 10 assessment criteria prior the application of OBM2 (five each for physical examination and clinical history stations. Assessment criteria for physical-examination-based and history-based stations were paired to create 5 new unified assessment criteria for the subsequent exam (see table 1). This grouping was conducted by three UNSW clinical examination experts, who together decided which criteria assessed similar skills and could therefore be paired together.

Data analysis compared the initial decision to the subsequent OSCE mark. The initial decision was used as the independent variable such that the results explore the predictive validity of the decisions. Using the original marks (prior the application of the OBM2) for the subsequent OSCE was important in order to avoid any unexpected unrelated impact the OBM2 might have had on the analysis. Therefore the analysis solely compared associations between decisions in the initial OSCEs and the (unmodified) subsequent OSCE marks.

The analysis was conducted using SPSS [22] starting with independent samples T-tests. Statistical significance was set at $p<0.05$. First, initial decisions within any initial assessment domain were compared to subsequent marks for any assessment criterion.

Further analysis explored the relationship between initial decisions per assessment domain and subsequent marks per assessment criterion. Accordingly, the association of initial decisions and subsequent assessment marks both within related domains, and across different domains can be determined. Cohen's d effect sizes were calculated for each individual factor [23].

Analysis of variance (ANOVA) tested between-subject effects to determine whether the station has a confounding effect on the association between initial decisions and subsequent assessment marks.

5. Results

Independent samples T-tests (see table 2 and figure 1) and ANOVA (see figure 2) demonstrate a statistically significant association between the initial decision and the subsequent OSCE performance (examination mark), one year later.

The T-test demonstrated that across 14 of all 15 comparisons, the subsequent OSCE marks related to initial pass decisions were significantly higher than subsequent OSCE marks related to initial fail decisions ($p<.05$) (see table 2 and figure 2). It is noted that small-medium effect sizes (Cohen's d=.223-.675) were identified across all the fourteen significant T-tests (see table 2).

Analysis comparing subsequent OSCE marks to initial decisions within each specific initial assessment domain demonstrated more specific links between initial decisions and subsequent marks (see table 2 and figure 1). With one exception; for every assessment domain, initial decisions have a predictive association with every subsequent assessment criterion. The exception is the relationship between decisions made for Initial physical examination, and subsequent history marks (p=.752, Cohen's d=.041) (see figure 1, section b).

Table 2: Independent Samples T-Test for the Association between Initial Decisions per Assessment Domain and Subsequent OSCE Marks per Assessment Criterion

Initial Assessment Domain	Subsequent Assessment Domain	Initial Fail		Initial Pass		t	df	Sig. (2tailed)	Mean Difference	Std. Error Difference	95% CI of Difference		Cohen's d Value
		N	Mean Subsequent Mark	N	Mean Subsequent Mark						Lower	Upper	
General Communication	Physical Examination	89	6.557	109	7.045	3.520	173.070	.001	-.488	.139	-.761	-.214	.500
	History	89	6.873	109	7.184	2.467	196.000	.014	-.311	.126	-.560	-.062	.351
	Communication	89	6.982	109	7.446	5.097	196.000	.000	-.464	.091	-.643	-.284	.725
	Case Interpretation	89	6.510	109	6.914	3.474	196.000	.001	-.404	.116	-.633	-.175	.494
	Case Summary	89	6.432	109	7.140	4.983	144.171	.000	-.708	.142	-.989	-.427	.708
Clinical Communication	Physical Examination	120	6.613	75	7.146	4.235	189.648	.000	-.533	.126	-.781	-.285	.607
	History	120	6.956	75	7.240	1.997	193.000	.047	-.285	.143	-.566	-.004	.286
	Communication	120	7.126	75	7.571	4.826	193.000	.000	-.445	.092	-.627	-.263	.691
	Case Interpretation	120	6.615	75	7.007	3.272	177.766	.001	-.392	.120	-.628	-.156	.469
	Case Summary	120	6.546	75	7.219	5.514	192.868	.000	-.674	.122	-.915	-.433	.790
Physical Examination	Physical Examination	113	6.601	123	7.073	3.885	234.000	.000	-.472	.121	-.711	-.233	.506
	History	113	7.204	123	7.238	.317	234.000	.752	-.034	.107	-.246	.178	.041
	Communication	113	7.295	123	7.520	2.982	234.000	.003	-.226	.076	-.375	-.077	.388
	Case Interpretation	113	6.654	123	6.972	2.994	234.000	.003	-.318	.106	-.528	-.109	.390
	Case Summary	113	6.687	123	7.176	4.288	199.674	.000	-.489	.114	-.714	-.264	.558

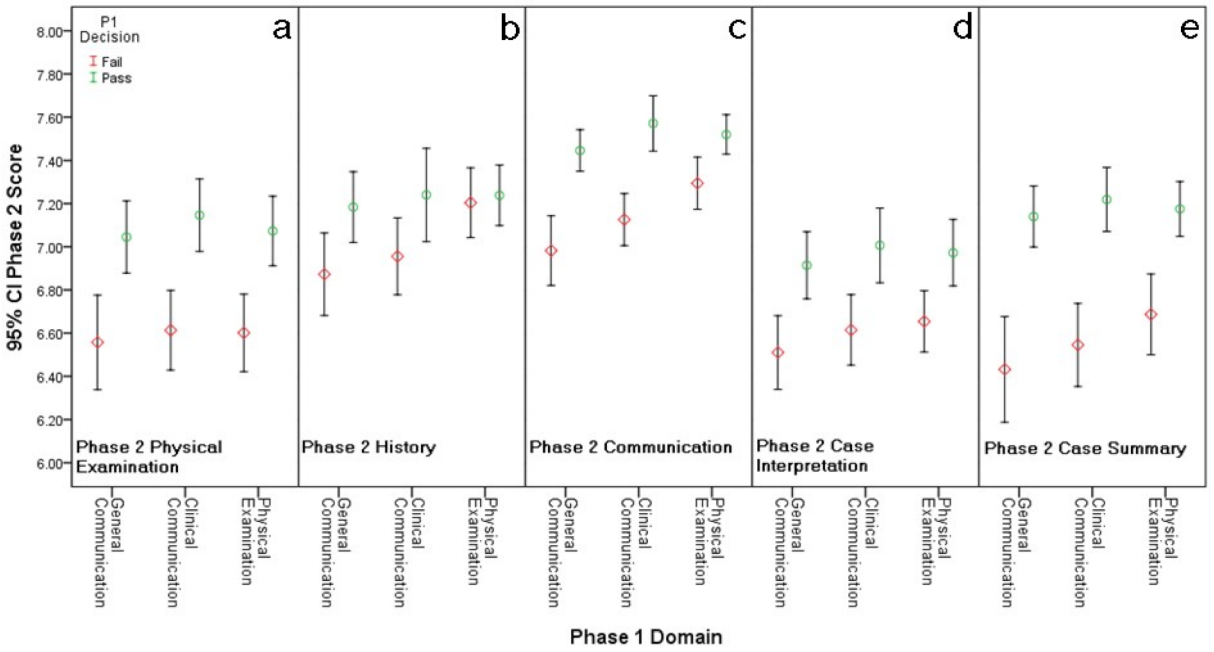

Figure 1: The Association between Initial Decisions per Assessment Domain and Subsequent OSCE Marks per Assessment Criterion

Effect sizes (Cohen's d) are larger when initial decisions per domain are compared to their related subsequent assessment criteria, than when the comparisons are made across less similar domains (see table 2). Both initial general communication and initial clinical communication have large effects on subsequent communication marks (Cohen's d=.725 and .691 respectively); furthermore, these two Initial domains have large effects on case summary (Cohen's d=.708 and .790 respectively) (see table 2). Similarly, initial decisions made for physical examination demonstrated a medium effect on subsequent physical examination marks (Cohen's d=.506). This also applies for initial physical examination decisions and subsequent case summary marks (Cohen's d=.558) (see table 2).

There is a similar statistically significant association in the ANOVA (see figure 2) for each comparison made between related Initial assessment domains and subsequent assessment criteria in independent samples T-tests.

Initial decisions made in the general communication domain were compared to marks for each subsequent assessment criterion. Similarly, initial decisions made in the general communication and physical examination domains were compared to subsequent marks per assessment criterion. This association again demonstrates that

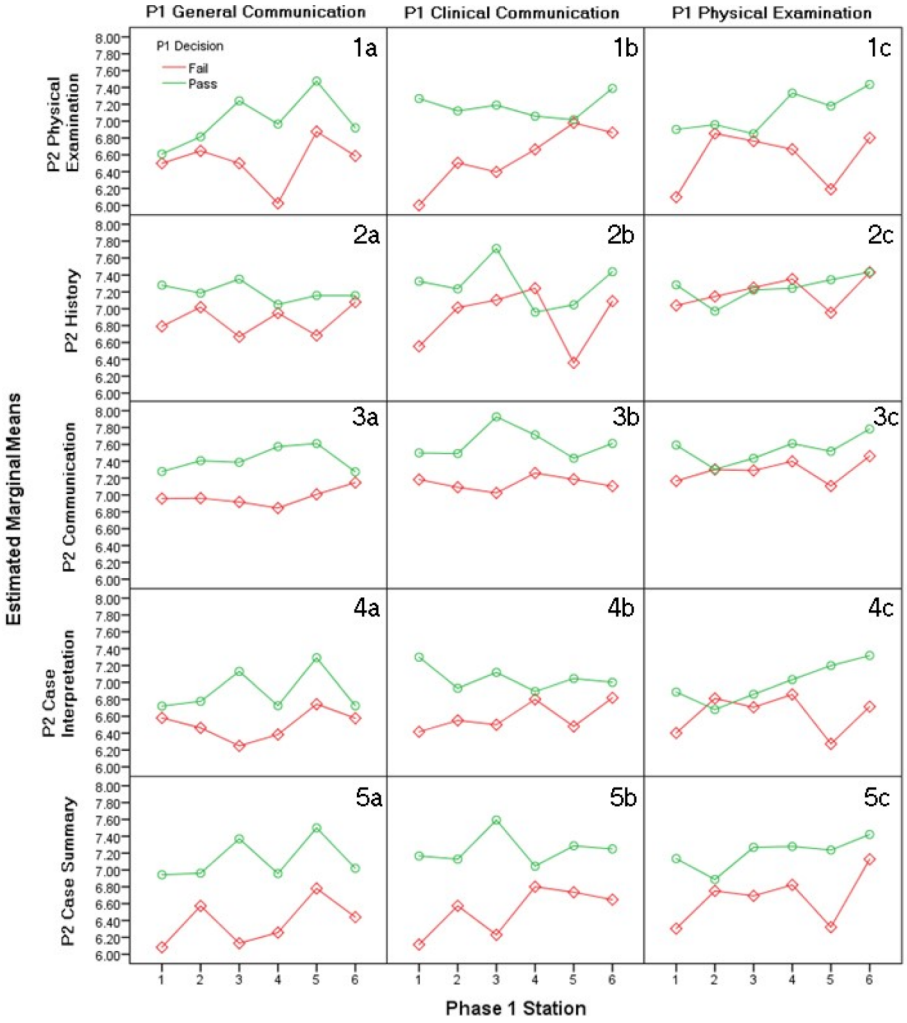

Figure 2: The Predictive Value of an Initial Decision per Assessment Domain for Subsequent OSCE Marks per Assessment Criterion by Station

initial pass decisions are associated with significantly higher ($p<0.05$) subsequent OSCE marks than initial fail decisions; particularly when related domains/criteria. Again, there is no significant association between subsequent history marks and initial decisions in the physical examination domain (see figure 2, section 2c).

Figure 2 demonstrates that there is a significant association between the initial decisions and subsequent OSCE scores. There are some outliers (see figure 2; sections 2b, 2c, 3c, 4c); however, an overall predictive association exists. Initial pass decisions resulted in consistently higher subsequent marks than initial fail decisions.

ANOVA determines that this predictive relationship is associated with the initial decision, independent of assessment station. These results indicate that the initial decisions were justified, as past grades should predict future performance, and have managed to do so based on these initial decisions.

6. Discussion

Initial decisions have a predictive association when applied to subsequent examinations within a cohort. This predictive validity is stronger within related initial assessment domains and subsequent assessment criteria than across less-related domains/criteria (see table 2, see figure 1 and figure 2).

A significant relationship between initial decisions and subsequent OSCE marks exists between initial general and clinical communication decisions, and subsequent history marks (see table 2; see figure 1, section b; see figure 2, section 2a-2b). Whereas, initial decisions in the physical examination domain have no significant association with subsequent history marks (see table 2; see figure 1, section b; see figure 2, section 2c). This is reasonable as the domains assess different skills, whereas communication and history assess similar skills.

Although all three initial assessment domains are significantly associated with subsequent communication marks; initial general and clinical communication decisions acted as substantially stronger predictors than initial physical examination (Cohen's d=.725, .691 and .388 respectively; see table 2; see figure 1, section c; see figure 2, sections 3a-3c). This demonstrates that although the predictive association exists across most domains, it remains strongest within the related domains.

Due to the requirement for case specificity in the phase 2 OSCEs, case interpretation relies on competent performance within a station to elicit information as well as underlying clinical knowledge to allow discovery and intellectual interpretation of case findings. This is demonstrated by the large effect size related to subsequent OSCE marks in case interpretation and case summary (see table 2). The UNSW Faculty of Medicine specifies that a good case summary relies on multiple factors assessed within the phase 2 OSCE including clear/concise general communication, appropriate clinical jargon, identification of significant case findings and suggestion of differential diagnoses [21].

Unmodified grades (borderline) are all identical and are reclassified according to the OBM2 decisions. There is no reason to expect such a predictive association unless decisions are valid. Repeated significant associations throughout different assessment domains/criteria (see figure 2) suggest that this predictability is not a random occurrence. These reclassified grades have a predictive association with future marks; such predictive associations are identified in literature [19]. The ability of decisions to mirror these expectations, especially within related assessment domains/criteria and less-so across unrelated domains/criteria enhances the validity of the decisions.

Multiple confounders, including the examiner, the examination site and the stations at which the student was examined may have an impact. Each of these is discussed below.

The UNSW Medicine Faculty uses various organisational strategies to mitigate judgement biases and avoid the occurrence of judgement errors. For the UNSW OSCEs, examiners are randomly selected and allocated to different examination sites. Assessors are rotated between different sites and external assessors are used [24]. Through this process, it is highly unlikely that the same student will be assessed by the same examiner in successive years.

UNSW data demonstrates that there is no significant difference in OSCE performance between different examination sites [24]. Furthermore, students are randomly allocated to an examination site for each OSCE, thus will not necessarily be assessed at the same site in consecutive years.

The phase 1 and phase 2 OSCEs are designed to satisfy different syllabi and assess different skills [21]. The OSCE stations at which the student is assessed will not be testing the same skill or clinical knowledge. Therefore, the station at which a student is assessed in the initial OSCE will not alter the association between initial decisions and subsequent OSCE marks. Additionally, ANOVA results establish that there is no significant association between the phase 1 station and phase 2 examination marks for any assessment domain/criterion.

After excluding each of these variables (examiner, examination site and examination stations), it is evident that most of the predictive nature is related to the decisions. This provides support for decisions to reclassify borderline grades to clear pass or clear fail grades. The validity of decisions has been asserted through a series of robust statistical tests. In conjunction with previous studies, this report provides further support for the validity of these decisions [7], [14], [17]. Consequently, these decisions resolve examiner uncertainty surrounding borderline scores. This may further increase the objectivity of pass/fail reclassification of borderline marks.

An important limitation is that the study used data from only one cohort of decisions at one university. The study would gain strength and reliability if the same tests were conducted for consecutive years' OSCE data from different cohorts and across different universities; as well as repetition on this cohort after completion of the third OSCE of the program, or comparison of the OBM2 to other standard setting methods, all of which may be explored in future studies.

7. Conclusion

Decisions have previously been shown to be effective, reliable, defensible and feasible. Previous studies have also suggested that decisions have acceptable validity. This is the first study to demonstrate the predictive validity of decisions, thus further supporting the validity of the decisions. These results may enhance examiners' confidence when making high-stakes decisions to reclassify borderline grades.

Further research may establish the OBM2's unknown limitations. A similar validation study could be repeated when phase 3 OSCE data is available for this cohort (year 2020), to determine whether similar predictive validity is maintained when tested across a third consecutive exam. Furthermore, the OBM2 could be tested within different settings and different examination styles.

Competing interests

The authors declare that they have no competing interests.

References

1. Rendel S, Foreman P, Freeman A. Licensing exams and judicial review: the closing of one door and opening of others? Br J Gen Pract. 2015;65(630):8-9. DOI: 10.3399/bjgp15X683029

2. Richard H, Sen GT, Jan V. The practical value of the standard error of measurement in borderline pass/fail decisions. Med Educ. 2008;42(8):810-815. DOI: 10.1111/j.1365-2923.2008.03103.x

3. Yudkowsky R, Tumuluru S, Casey P, Herlich N, Ledonne C. A Patient Safety Approach to Setting Pass/Fail Standards for Basic Procedural Skills Checklists. Simul Healthc. 2014;9(5):277-282. DOI: 10.1097/SIH.0000000000000044

4. Cizek GJ, Bunch MB. Standard setting: A guide to establishing and evaluating performance standards on tests. Thousand Oaks (CA): SAGE Publications Ltd; 2006.

5. Ben-David MF. AMEE Guide No. 18: Standard setting in student assessment. Med Teach. 2000;22(2):120-130. DOI: 10.1080/01421590078526
6. Phillips G. Technical Issues in Large-Scale Performance Assessment. Washington: U.S. Department of Education; 1996.
7. Shulruf B, Coombes L, Damodaran A, Freeman A, Jones P, Lieberman S, Poole P, Rhee J, Wilkinson T, Harris P. Cut-scores revisited: feasibility of a new method for group standard setting. BMC Med Educ. 2018;18(1):126. DOI: 10.1186/s12909-018-1238-7
8. Shulruf B, Wilkinson T, Weller J, Jones P, Poole P. Insights into the Angoff method: results from a simulation study. BMC Med Educ. 2016;16:134. DOI: 10.1186/s12909-016-0656-7
9. Hurtz GM, Hertz NR. How Many Raters Should be Used for Establishing Cutoff Scores with the Angoff Method? A Generalizability Theory Study. Educ Psychol Measurement. 1999;59(6):885-897. DOI: 10.1177/00131649921970233
10. Shulruf B, Turner R, Poole P, Wilkinson T. The Objective Borderline method (OBM): a probability-based model for setting up an objective pass/fail cut-off score for borderline grades in medical education programmes. Adv Health Sci Educ Theory Pract. 2013;18(2):231-144. DOI: 10.1007/s10459-012-9367-y
11. Wood T, Humphrey-Murto S, Norman G. Standard Setting in a Small Scale OSCE: A Comparison of the Modified Borderline-Group Method and the Borderline Regression Method. Adv Health Sci Educ Theory Pract. 2006;11(2):115-122. DOI: 10.1007/s10459-005-7853-1
12. Behuniak P, Archambault F, Gable R. Angoff and Nedelsky Standard Setting Procedures: Implications for the Validity of Proficiency Test Score Interpretation. Educ Psychol Measurement. 1982;42(1):247-255. DOI: 10.1177/0013164482421031
13. Poggio JP. An Empirical Investigation of the Angoff, Ebel and Nedelsky Standard Setting Methods. In: 65th Annual Meeting of the American Educational Research Association; 1981 Apr 13-17; Los Angeles, CA, United States. Zugänglich unter/available from: https://eric.ed.gov/?id=ED205552
14. Shulruf B, Poole P, Jones P, Wilkinson T. The Objective Borderline Method: a probabilistic method for standard setting. Ass Eval High Educ. 2015;40(3):420-438. DOI: 10.1080/02602938.2014.918088
15. Shulruf B, Adelstein BA, Damodaran A, Harris P, Kennedy S, O'Sullivan A, Taylor S. Borderline grades in high stakes clinical examinations: resolving examiner uncertainty. BMC Med Educ. 2018;18(1):272. DOI: 10.1186/s12909-018-1382-0
16. Shulruf B, Damodaran A, Jones P, Kennedy S, Mangos G, O'Sullivan A, Rhee J, Tayler S, Velan G, Harris P. Enhancing the defensibility of examiners' marks in high stake OSCEs. BMC Med Educ. 2018;18(1):10. DOI: 10.1186/s12909-017-1112-z
17. Shulruf B, Booth R, Baker H, Bagg W, Barrow M. Using the Objective Borderline Method (OBM) to support Board of Examiners' decisions in a medical programme. J Furth High Educ. 2017;41(3):425-434. DOI: 10.1080/0309877X.2015.1117603
18. Garson D. Validity and Reliabiity. North Carolina: Statistical Publishing Associates; 2016.
19. Poole P, Shulruf B, Rudland J, Wilkinson T. Comparison of UMAT scores and GPA in prediction of performance in medical school: a national study. Med Educ. 2012;46(2):163-171. DOI: 10.1111/j.1365-2923.2011.04078.x
20. University of New South Wales, Faculty of Medicine. Phase 1 / Graduate Entry Clinical Skills Student Guide 2018. Kensington: The University of New South Wales; 2018.
21. University of New South Wales, Faculty of Medicine. Phase 2 Clinical SKills Guide 2018. Kensington: The University of New South Wales; 2018.
22. IBM Corporation. IBM SPSS Statistics for Windows. 24 ed. Armonk, NY: IBM Corporation; 2016.
23. Wilson D. Practical Meta-Analysis Effect Size Calculator. Fairfax: George Mason University; 2018.
24. Medical School Accreditation Committee. Accreditation of University of New South Wales Faculty of Medicine. Kingston: Australia Medical Council Limited; 2018.

Corresponding author:
Boaz Shulruf
University of New South Wales, Office of Medical Education, Sydney, Australia
b.shulruf@unsw.edu.au

Please cite as
Klein Nulend R, Harris P, Shulruf B. Predictive validity of a tool to resolve borderline grades in OSCEs. GMS J Med Educ. 2020;37(3):Doc31. DOI: 10.3205/zma001324, URN: urn:nbn:de:0183-zma0013243

This article is freely available from
https://www.egms.de/en/journals/zma/2020-37/zma001324.shtml

Received: *2019-03-18*
Revised: *2019-11-19*
Accepted: *2020-01-07*
Published: *2020-04-15*

OPEN ACCESS

This is the German version.
The English version starts at p. 1.

Vorhersagevalidität eines Instruments zur Beseitigung von Grenznoten bei OSCE-Prüfungen

Rowan Klein Nulend[1]
Peter Harris[1]
Boaz Shulruf[1]

1 University of New South Wales, Office of Medical Education, Sydney, Australien

Zusammenfassung

Es liegen keine eindeutigen Belege dafür vor, welches Standard-Setting-Verfahren die höchste Validität für Entscheidungen über das Bestehen bzw. Nichtbestehen von Prüfungen ergibt. Die Objective Borderline Method 2 (OBM2) ist ein Instrument zur Entscheidungsunterstützung für die Neueinstufung von Grenznoten als eindeutige Bestehens- oder Nichtbestehensnoten. So können Unsicherheiten der Prüfer bei folgenreichen Entscheidungen über das Bestehen bzw. Nichtbestehen beseitigt werden.

In dieser Studie wurde die Vorhersagevalidität von OBM2-basierten Entscheidungen über das Bestehen bzw. Nichtbestehen unter Verwendung der Ergebnisse der Objective Structured Clinical Examination (OSCE) in aufeinanderfolgenden Jahren innerhalb einer Kohorte von Medizinstudenten (N 71) an der University of New South Wales, Australien, geprüft. OBM2-basierte Entscheidungen in einer OSCE-Prüfung (N=687) wurden mit den in einer darauffolgenden OSCE-Prüfung erhaltenen Noten anhand von t-Tests für unabhängige Stichproben und einer Varianzanalyse (ANOVA) verglichen. Der Umfang des Zusammenhangs zwischen diesen beiden Variablen bestimmt die Vorhersagevalidität von OBM2-basierten Entscheidungen, vorausgesetzt, die vorherigen Noten der Studenten lassen Vorhersagen zur zukünftigen Leistung zu. Es wurde gezeigt, dass durch OBM2-basierte Entscheidungen bei einer ersten OSCE-Prüfung statistisch signifikante Vorhersagen für die nachfolgenden OSCE-Noten (p=0,005) getroffen werden können. In den Fällen, in denen die Noten aus der ersten Prüfung als Bestehensnote neueingestuft wurden, waren die nachfolgenden OSCE-Noten signifikant besser als in den Fällen, in denen die Noten aus der ersten Prüfung als Nichtbestehensnote neueingestuft wurden. Ein stärkerer Zusammenhang wurde für verwandte Bewertungsdomänen/-kriterien im Vergleich zu nicht verwandten Domänen/Kriterien gefunden (Cohens d=0,469 versus Cohens d=0,388).

Der gezeigte prädiktive Zusammenhang der OBM2-basierten Entscheidungen über Prüfungen hinweg stützt die Vorhersagevalidität der OBM2. Sie wird daher als eine vielversprechende Methode zur Beseitigung von Unsicherheiten der Prüfer bei Entscheidungen über das Bestehen bzw. Nichtbestehen von OSCE-Prüfungen betrachtet.

Schlüsselwörter: OCSE, Grenznoten, Bewertung, Medizinstudenten

1. Einleitung

Auf Bewertungsstrategien basierende Entscheidungen im Rahmen eines Medizinstudiengangs müssen belastbar sein [1], [2], [3]. Subjektivität mindert die Belastbarkeit von Prüfungen; um die Objektivität bei OSCE-Prüfungen zu erhöhen, wird daher meist ein Standard-Setting-Verfahren angewendet [4]. Standard-Setting-Verfahren dienen zur Bestimmung von Cut-Scores, die dem Mindestmaß an erforderlicher Kompetenz/Leistung bei einer Bewertungsaufgabe entsprechen [4], [5].

Es stehen viele Standard-Setting-Verfahren zur Verfügung; alle in der aktuellen Literatur untersuchten Verfahren weisen jedoch subjektive Aspekte und Ungenauigkeiten sowie uneindeutige Daten zu deren Effizienz auf [6], [7]. In den meisten Verfahren ist eine Beurteilung durch Experten/Prüfer vorgesehen. Obwohl diese Beurteilungen von Experten auf dem jeweiligen Gebiet erfolgen, ist eine vollständige Objektivität in diesen Fällen niemals möglich [6], [8], [9]. Da es für das Standard-Setting keinen Goldstandard gibt, ist hierbei die Validierung eines Standard-Setting-Verfahrens die größte Herausforderung [8]. In früheren Studien wurde gezeigt, dass bei der Anwendung zweier oder mehrerer Standard-Setting-Verfahren auf

denselben Datensatz jedes Verfahren einen anderen Cut-Score ergibt [10], [11], [12].
Weitere Probleme sind die Definition des Grenzwerts, meist auch bezeichnet als „minimal kompetenter" Student, und die Variabilität der Expertenmeinung in dieser Domäne [13]. Ein Grenzergebnis liegt vor, wenn der Prüfer die abgelieferte Leistung nicht eindeutig als „bestanden" oder „nicht bestanden" einordnen kann. Dieser Fall tritt ein, wenn die Leistung des Studenten nah am erwarteten Cut-Score liegt, der zwischen Bestehens- und Nichtbestehensnoten unterscheidet [4].
Zur Lösung dieses Problems wurde die Objective Borderline Method (OBM) eingeführt [10]. Die OBM ist ein Standard-Setting-Verfahren, bei dem Grenznoten nachträglich entweder als Bestehens- oder Nichtbestehensnoten eingestuft werden. Die Grundlage dafür bildet das Verhältnis der Prüfungsteilnehmer, die bestanden, nicht bestanden oder Grenznoten erhalten haben [14]. Dieses Modell basiert auf Wahrscheinlichkeit unter Verwendung des Verhältnisses von Bestehens-/Nichtbestehens-/Grenznoten. Bei den meisten Standard-Setting-Verfahren wird ein Cut-Score anhand von Expertenmeinung oder statistischen Verfahren zugewiesen, wie beispielsweise bei der Angoff-Methode bzw. der Borderline-Regression-Methode [10].
Nach der Einführung der OBM wurde die Objective Borderline Method 2 (OBM2) entwickelt. Die OBM2 ist kein Standard-Setting-Verfahren, da hierbei kein Cut-Score ermittelt wird. Sie ist vielmehr ein Instrument zur Entscheidungsunterstützung für die Neueinstufung von Grenznoten. Hierbei werden nur zwei Größen verwendet, um die Neueinstufung der Grenznoten als Bestehens- oder Nichtbestehensnoten auf Einzelfallbasis vorzunehmen: die Fähigkeiten des Prüfungsteilnehmers und die Aufgabenschwierigkeit, die anhand aller Noten einer Prüfung ermittelt wird. Es wurde festgestellt, dass die OBM2 für standardisierte Prüfungen der klinischen Kompetenz anwendbar ist, um Entscheidungen über das Bestehen bzw. Nichtbestehen bei Grenznoten zu treffen [15].
Die OBM2 ist eine wahrscheinlichkeitsbasierte Methode, um die Grenznote eines Prüfungsteilnehmers durch eine Bestehens- oder Nichtbestehensnote für jede Aufgabe zu ersetzen [16], [17]. Somit kann ein Prüfungsteilnehmer beliebig viele Grenznoten zwischen Null und der Gesamtzahl der Aufgaben der Prüfung erhalten (in der aktuellen Studie liegt die Anzahl zwischen 0 und 54 pro Student). Eine Grenznote wird vom Prüfer dann vergeben, wenn er die Ausführung des Prüfungsteilnehmers nicht eindeutig als „bestanden" oder „nicht bestanden" einordnen kann [16], [17]. Die Neueinstufung von Grenznoten entweder als Bestehens- oder Nichtbestehensnoten erfolgt anhand des Verhältnisses der Bestehens- (p), Grenz- (b) und Nichtbestehensnoten (f) der Studenten unter Verwendung der folgenden Formel: „OBM-Index=(p/[b+p])×(b/[f+b])" [16]. Der OBM-Index wird zweimal berechnet: für die vom Studenten erhaltenen Noten für alle Aufgaben zur Bestimmung der „Fähigkeit des Studenten" und für die von allen Studenten erhaltenen Noten für jede Aufgabe zur Bestimmung der „Aufgabenschwierigkeit". Somit ergeben sich für jede Grenznote zwei OBM-Indizes. Anschließend werden pro Grenznote die beiden OBM-Indizes verglichen. Wenn „Fähigkeit des Studenten"≥„Aufgabenschwierigkeit", wird die Grenznote als Bestehensnote neueingestuft. Wenn „Fähigkeit des Studenten"<„Aufgabenschwierigkeit", wird die Grenznote als Nichtbestehensnote neueingestuft. Die technischen Einzelheiten der OBM2 wurden in früheren Untersuchungen gezeigt [16].
In der Lehre ist die Vorhersagevalidität ein wichtiger Bestandteil der Kriteriumsvalidität, da die Vorhersage der zukünftigen Leistung ein wesentliches Ziel bei Prüfungen ist [18]. Der aktuellen Literatur zufolge können anhand der vorherigen Noten des Studenten Vorhersagen zur zukünftigen Leistung getroffen werden [19]. Wenn die OBM2 diese Erwartung innerhalb einer Gruppe von Studenten mit derselben Note (Grenznote) und anschließender Neueinstufung als Bestehens- oder Nichtbestehensnote widerspiegeln könnte, würde dies die Validität der OBM2 als Instrument zur Neueinstufung von Grenznoten als „eindeutig bestanden" oder „eindeutig nicht bestanden" erhöhen. Dies trifft dann zu, wenn ein Student mit einer Grenznote anhand der OBM2-basierten Entscheidung in eine Gruppe einordnet wird, die der tatsächlichen zukünftigen Leistung entspricht, die anhand der vorherigen Noten des Studenten erwartet wurde.
In früheren Studien wurden die OBM2 erklärt und die Belastbarkeit, die Praktikabilität, der Einfluss auf die OSCE-Ergebnisse und die Validität dieses Instruments beurteilt [14], [16], [17]. Für diese Studien wurden jedoch Augenblicksdaten verwendet, die keine Hinweise zur Vorhersagevalidität von OBM2-basierten Entscheidungen über das Bestehen/Nichtbestehen lieferten [10], [14], [16].

2. Ziel

Die Studie zielte darauf ab, zu bestimmen, inwieweit anhand von OBM2-basierten Entscheidungen die zukünftigen Leistungen vorhergesagt werden können. Dadurch kann die Vorhersagevalidität von OBM2-basierten Entscheidungen über das Bestehen/Nichtbestehen bestimmt werden. Zum Erreichen des Ziels wurde die folgende Forschungsfrage formuliert: Wie stark ist der Zusammenhang zwischen OBM2-basierten Entscheidungen in einer OSCE-Prüfung und den in der OSCE-Prüfung des darauffolgenden Jahres erhaltenen Noten?

3. Studienaufbau

In dieser Studie wurden Daten aus an der University of New South Wales (UNSW) in Sydney, Australien, durchgeführten OSCE-Prüfungen verwendet. Das Medizinstudium an der UNSW dauert sechs Jahre bis zum ersten Abschluss. Jeweils im zweiten, im dritten und im sechsten Jahr ist eine OSCE-Prüfung vorgesehen [20]. In dieser Studie wurden Daten aus den OSCE-Prüfungen im zweiten Jahr (als „erste" OSCE-Prüfung bezeichnet) und im dritten

Jahr (als „nachfolgende“ OSCE-Prüfung bezeichnet) derselben Kohorte in zwei aufeinanderfolgenden Jahren (2016/2017) verwendet. In den ersten beiden Jahren des Medizinstudiums an der UNSW wird vorranging Theorie vermittelt. Dabei sind die wöchentlich wechselnden zweistündigen Veranstaltungen zum Erlernen klinischer Fähigkeiten auf dem Campus und im Krankenhaus der einzige klinische Praxisunterricht für die Studenten. Im gesamten dritten Jahr sind die Studenten täglich in einem zugewiesenen Krankenhaus tätig, wobei sie wesentlich mehr klinische Erfahrung sammeln [17], [18].

In der ersten Prüfung werden die Studenten (N=271) in drei Domänen geprüft: allgemeine Kommunikation, klinische Kommunikation und körperliche Untersuchung. Diese Domänen sind in jeweils neun spezifische Bewertungskriterien innerhalb der Benotungsrubrik unterteilt. So kann ein Student bis zu neun Grenzergebnisse pro OSCE-Station erzielen. Die Kohorte war auf vier Standorte verteilt [21]. In der nachfolgenden Prüfung (257 Studenten) werden leicht unterschiedliche Bewertungskriterien verwendet (siehe Tabelle 1) [21]. Diese Prüfung findet an neun Standorten statt.

Die erste und die nachfolgende OSCE-Prüfung bestehen aus sechs verschiedenen Stationen mit unterschiedlichen Fällen und Prüfern [21]. Jede Station wird von einem Prüfer bewertet. Die Prüfer sind sowohl externe Personen als auch Universitätsangehörige. In der ersten OSCE-Prüfung sind 15 Minuten pro Station vorgesehen. Hier stehen die Bewertung der klinischen Fähigkeiten wie die allgemeine Kommunikation, die klinische Kommunikation und die körperliche Untersuchung im Vordergrund [21]. In der nachfolgenden OSCE-Prüfung sind 10 Minuten pro Station vorgesehen. Hier wird neben diesen klinischen Fähigkeiten auch die Fallspezifität bewertet. Somit ist ein umfangreiches klinisches Wissen für eine gute Leistung in dieser Prüfung notwendig [21]. Die Kriterien der nachfolgenden OSCE-Prüfung haben Äquivalente in den drei Domänen der ersten OSCE-Prüfung, wodurch Vergleiche möglich sind. Sowohl für die erste als auch die nachfolgende OSCE-Prüfung ist bei Nichtbestehen ein Zweitversuch möglich. Die Prüfer der nachfolgenden OSCE-Prüfung kannten die Noten der Studenten in der ersten OSCE-Prüfung nicht. Die Studie umfasste Daten von 271 Studenten, die im Jahr 2016 an der OSCE-Prüfung des zweiten Jahres teilnahmen. Diese OSCE-Prüfung umfasst sechs Stationen. An jeder dieser Stationen wird der Student anhand von neun Bewertungskriterien bewertet. Daraus ergeben sich so 54 Noten pro Student für die OSCE-Prüfung des zweiten Jahres. Bei jedem Bewertungskriterium liegt das Augenmerk auf einer der drei Domänen „allgemeine Kommunikation“, „klinische Kommunikation“ oder „körperliche Untersuchung“. Insgesamt wurden in der OSCE-Prüfung des zweiten Jahres 14.634 Noten vergeben (f=83 [0,6%]; b=687 [4,7%]; p=13864 [94,7%], die p-Note umfasst die Noten für „bestanden“ und „Prädikatsnote“). Nach Anwendung der OBM2, wodurch die Grenznoten durch Bestehens- oder Nichtbestehensnoten ersetzt wurden, wurden die Noten je Domäne summiert (gemittelt) und als solche erfasst. In dieser Studie wurden jedoch nur die 687 Grenznoten untersucht, denn nur diese wurden in Bestehens- oder Nichtbestehensnoten geändert.

4. Methoden

Im Folgenden werden „OMB2-basierte Entscheidungen zur Neueinstufung von Grenznoten als eindeutige Bestehens- oder Nichtbestehensnoten“ als „Entscheidungen“ bezeichnet.

Ein Datensatz umfasste alle Grenznoten der ersten Prüfung, für die Entscheidungen getroffen wurden (N=687); der zweite Datensatz umfasste alle Noten der nachfolgenden Prüfung, die mit jeder Entscheidung in der ersten Prüfung korrelierten. Im Falle von 58 der 687 Entscheidungen über Grenznoten in der ersten Prüfung (14 Studenten) traten die betreffenden Studenten die nachfolgende OSCE-Prüfung im darauffolgenden Jahr nicht an. Die Einträge für die nachfolgende Prüfung waren somit nicht vollständig und blieben daher in der Analyse unberücksichtigt. Folglich wurden 629 Entscheidungsgruppen (257 Studenten) analysiert. In der ersten Prüfung können die Studenten maximal neun Grenznoten pro OSCE-Station erhalten, da sie jeweils anhand von neun Kriterien pro Station bewertet werden.

Die Daten für die nachfolgende Prüfung umfassten die ursprünglichen Noten für zehn Bewertungskriterien vor der Anwendung der OBM2 (jeweils fünf für die Stationen für die körperliche Untersuchung und die Stationen für die Krankengeschichte). Die Bewertungskriterien für die Stationen für die körperliche Untersuchung und die Stationen für die Krankengeschichte wurden paarweise zusammengefasst, um fünf neue einheitliche Bewertungskriterien für die nachfolgende Prüfung zu erstellen (siehe Tabelle 1). Diese Zusammenfassung nahmen drei Experten für Prüfungen auf dem Gebiet der klinischen Fähigkeiten von der UNSW vor. Sie entschieden gemeinsam, anhand welcher Kriterien ähnliche Fähigkeiten bewertet werden. Diese Kriterien wurden dann paarweise zusammengefasst.

In einer Datenanalyse wurde die für die erste Prüfung getroffene Entscheidung mit der in der nachfolgenden Prüfung erreichten Note verglichen. Die für die erste Prüfung getroffene Entscheidung wurde als unabhängige Variable so verwendet, dass die Ergebnisse die Vorhersagevalidität der Entscheidung herausstellt. Die Verwendung der ursprünglichen Noten (vor der Anwendung der OBM2) für die nachfolgende OSCE-Prüfung war wichtig, um einen unerwarteten zusammenhanglosen Einfluss der OBM2 auf die Analyse auszuschließen. Daher bestand die Analyse lediglich aus dem Vergleich der Zusammenhänge der Entscheidungen in den ersten OSCE-Prüfungen mit den in den nachfolgenden OSCE-Prüfungen erhaltenen (unveränderten) Noten.

Die Analyse wurde mithilfe von SPSS [22] durchgeführt. Es wurde mit t-Tests für unabhängige Stichproben begonnen. Statistische Signifikanz wurde bei $p<0,05$ angenommen. Zunächst wurden die Entscheidungen in der ersten

Tabelle 1: Von klinischen Experten erstellte Gesamtbewertungskriterien für die nachfolgende Prüfung

Bewertungskriterium Station für Krankengeschichte	Bewertungskriterium Station für körperliche Untersuchung	Einheitliches Kriterium
Fachlich kompetente körperliche Untersuchung oder Fähigkeit	Erkennen körperlicher Symptome	Körperliche Untersuchung
Eruieren der relevanten Krankengeschichte	Zusammentragen der relevanten psychosozialen, medizinischen und familiären Vorgeschichte	Vorgeschichte
Aufmerksames Zuhören, Einbeziehen des Patienten, Aufrechterhalten von Respekt	Einbeziehen des Patienten, Aufrechterhalten von Respekt	Kommunikation
Interpretieren der Krankengeschichte und Vorstellen	Interpretieren des Patientenfalls	Fallinterpretation
Zusammenfassen der Vorgeschichte für den Prüfer	Zusammenfassen der Fallergebnisse	Fallzusammenfassung

Prüfung innerhalb jeder Bewertungsdomäne dieser ersten Prüfung mit den in der nachfolgenden Prüfung erhaltenen Noten für jedes Bewertungskriterium verglichen.
In weiteren Analysen wurde der Zusammenhang zwischen den Entscheidungen in der ersten Prüfung pro Bewertungsdomäne und den in der nachfolgenden Prüfung erhaltenen Noten pro Bewertungskriterium untersucht. Dementsprechend konnte der Zusammenhang der Entscheidungen in der ersten Prüfung und den in der nachfolgenden Prüfung erhaltenen Noten innerhalb verwandter Domänen sowie über verschiedene Domänen hinweg bestimmt werden. Für jeden einzelnen Faktor wurde die Effektstärke Cohens d berechnet [23].
Mithilfe der Varianzanalyse (ANOVA) wurden Tests der Zwischensubjekteffekte durchgeführt, um zu bestimmen, ob die Station den Zusammenhang zwischen den Entscheidungen in der ersten Prüfung und den in der nachfolgenden Prüfung erhaltenen Noten verzerrt.

5. Ergebnisse

Die t-Tests für unabhängige Stichproben (siehe Tabelle 2 und Abbildung 1) und die ANOVA (siehe Abbildung 2) zeigen einen statistisch signifikanten Zusammenhang zwischen der Entscheidung in der ersten Prüfung und der Leistung in der nachfolgenden OSCE-Prüfung (Prüfungsnote) ein Jahr später.
Der t-Test zeigte, dass bei 14 der insgesamt 15 Vergleiche die in der nachfolgenden OSCE-Prüfung erhaltenen Noten, die den Entscheidungen für Bestehensnoten zugehörig sind, signifikant besser waren als die in der nachfolgenden OSCE-Prüfung erhaltenen Noten, die den Entscheidungen für Nichtbestehensnoten zugehörig sind ($p<0,05$) (siehe Tabelle 2 und Abbildung 2). Es ist anzumerken, dass eine geringe bis mittlere Effektstärke (Cohens d=0,223–0,675) bei allen 14 signifikanten t-Tests gefunden wurde (siehe Tabelle 2).
Analysen zum Vergleich der in der nachfolgenden OSCE-Prüfung erhaltenen Noten mit den Entscheidungen in der ersten Prüfung innerhalb jeder der spezifischen Bewertungsdomänen dieser ersten Prüfung zeigten noch spezifischere Verbindungen zwischen den Entscheidungen in der ersten Prüfung und den in der nachfolgenden Prüfung erhaltenen Noten (siehe Tabelle 2 und Abbildung 1). Die Entscheidungen in der ersten Prüfung weisen mit einer Ausnahme für jede Bewertungsdomäne einen prädiktiven Zusammenhang mit jedem in der nachfolgenden Prüfung angewendeten Bewertungskriterium auf. Die Ausnahme ist der Zusammenhang zwischen den für „körperliche Untersuchung“ in der ersten Prüfung getroffenen Entscheidungen und den Noten für „Vorgeschichte“ in der nachfolgenden Prüfung ($p=0,752$, Cohens d=0,41) (siehe Abbildung 1, Feld b).
Die Effektstärken (Cohens d) sind höher, wenn die in der ersten Prüfung getroffenen Entscheidungen pro Domäne mit ihren in der nachfolgenden Prüfung angewendeten verwandten Bewertungskriterien verglichen werden, als wenn die Vergleiche über weniger verwandte Domänen erfolgen (siehe Tabelle 2). Sowohl die „allgemeine Kommunikation“ als auch die „klinische Kommunikation“ in der ersten Prüfung haben große Effekte auf die Noten im Bereich „Kommunikation“ in der nachfolgenden Prüfung (Cohens d=0,725 bzw. 0,691); darüber hinaus haben diese zwei Domänen aus der ersten Prüfung große Effekte auf die „Fallzusammenfassung“ (Cohens d=0,708 bzw. 0,790) (siehe Tabelle 2). Gleichermaßen zeigten die in der ersten Prüfung getroffenen Entscheidungen in der Domäne „körperliche Untersuchung“ einen mittleren Effekt auf die in der nachfolgenden Prüfung erhaltenen Noten für „körperliche Untersuchung“(Cohens d=0,506). Dies trifft auch auf die Entscheidungen in den ersten Prüfung in der Domäne „körperliche Untersuchung“ und die in der nachfolgenden Prüfung erhaltenen Noten für „Fallzusammenfassung“ zu (Cohens d=0,558) (siehe Tabelle 2).
In der ANOVA zeigt sich ein ähnlicher statistisch signifikanter Zusammenhang (siehe Abbildung 2) für jeden Vergleich zwischen verwandten Bewertungsdomänen der

Tabelle 2: t-Test für unabhängige Stichproben für den Zusammenhang zwischen den Entscheidungen in der ersten Prüfung pro Bewertungsdomäne und den in der nachfolgenden OSCE-Prüfung erhaltenen Noten pro Bewertungskriterium

Bewertungsdomäne der ersten Prüfung	Bewertungsdomäne der nachfolgenden Prüfung	1.Prfg. n.best. N	1.Prfg. n.best. Mittl. Note 2. Prfg.	1.Prfg. best. N	1.Prfg. best. Mittl. Note 2. Prfg.	t	df	Sig. (2seitig)	Mittl. Differenz	Std. fehlerdifferenz	95%-KI d. Differenz Untere	95%-KI d. Differenz Obere	Cohens d
Allgemeine Kommunikation	Körperl. Untersuchung	89	6,557	109	7,045	3,520	173,070	,001	–,488	,139	–,761	–,214	,500
	Vorgeschichte	89	6,873	109	7,184	2,467	196,000	,014	–,311	,126	–,560	–,062	,351
	Kommunikation	89	6,982	109	7,446	5,097	196,000	,000	–,464	,091	–,643	–,284	,725
	Fallinterpretation	89	6,510	109	6,914	3,474	196,000	,001	–,404	,116	–,633	–,175	,494
	Fallzusammenfassung	89	6,432	109	7,140	4,983	144,171	,000	–,708	,142	–,989	–,427	,708
Klinische Kommunikation	Körperl. Untersuchung	120	6,613	75	7,146	4,235	189,648	,000	–,533	,126	–,781	–,285	,607
	Vorgeschichte	120	6,956	75	7,240	1,997	193,000	,047	–,285	,143	–,566	–,004	,286
	Kommunikation	120	7,126	75	7,571	4,826	193,000	,000	–,445	,092	–,627	–,263	,691
	Fallinterpretation	120	6,615	75	7,007	3,272	177,766	,001	–,392	,120	–,628	–,156	,469
	Fallzusammenfassung	120	6,546	75	7,219	5,514	192,868	,000	–,674	,122	–,915	–,433	,790
Körperliche Untersuchung	Körperl. Untersuchung	113	6,601	123	7,073	3,885	234,000	,000	–,472	,121	–,711	–,233	,506
	Vorgeschichte	113	7,204	123	7,238	,317	234,000	,752	–,034	,107	–,246	,178	,041
	Kommunikation	113	7,295	123	7,520	2,982	234,000	,003	–,226	,076	–,375	–,077	,388
	Fallinterpretation	113	6,654	123	6,972	2,994	234,000	,003	–,318	,106	–,528	–,109	,390
	Fallzusammenfassung	113	6,687	123	7,176	4,288	199,674	,000	–,489	,114	–,714	–,264	,558

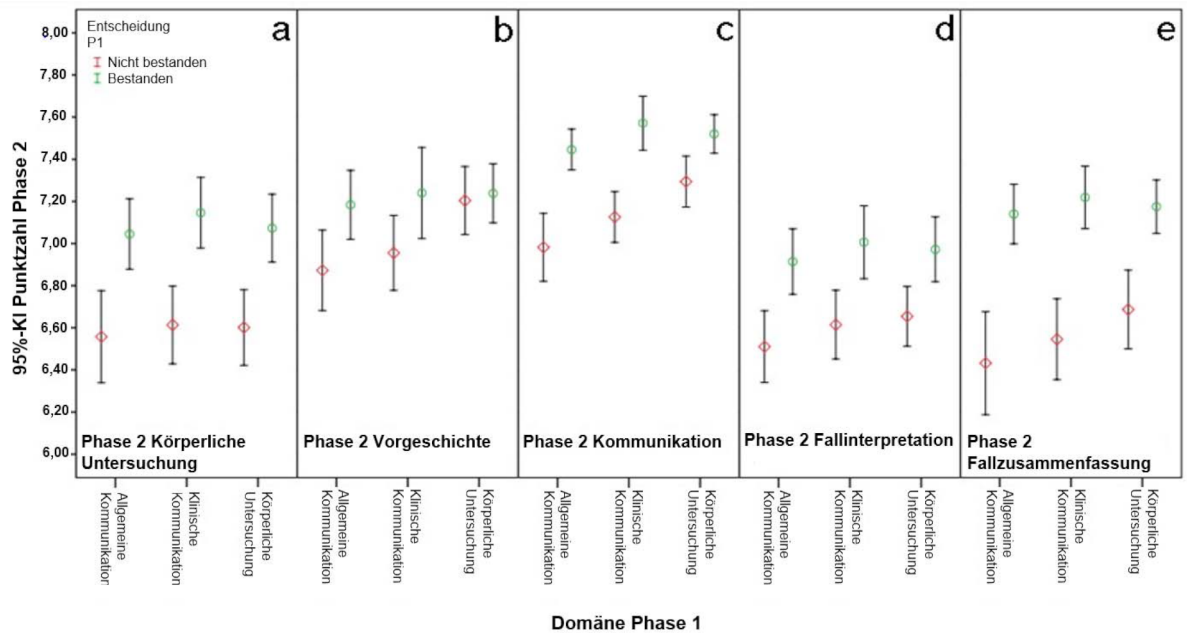

Abbildung 1: Der Zusammenhang zwischen den Entscheidungen in der ersten Prüfung pro Bewertungsdomäne und den in der nachfolgenden OSCE-Prüfung erhaltenen Noten pro Bewertungskriterium

ersten Prüfung und Bewertungskriterien der nachfolgenden Prüfung in t-Tests für unabhängige Stichproben.

Entscheidungen in der ersten Prüfung in der Domäne „allgemeine Kommunikation“ wurden mit den Noten für jedes Bewertungskriterium der nachfolgenden Prüfung verglichen. Gleichermaßen wurden die Entscheidungen in der ersten Prüfung in den Domänen „allgemeine Kommunikation“ und „körperliche Untersuchung“ mit den Noten für jedes Bewertungskriterium der nachfolgenden Prüfung verglichen. Dieser Zusammenhang zeigt erneut, dass Entscheidungen für Bestehensnoten in der ersten Prüfung mit signifikant besseren ($p<0,05$) Noten in der nachfolgenden OSCE-Prüfung in Zusammenhang stehen als Entscheidungen für Nichtbestehensnoten in der ersten Prüfung; dies trifft vor allem bei verwandten Domänen/Kriterien zu. Wieder besteht kein statistischer Zusammenhang zwischen Noten in der Domäne „Vorgeschichte“ in der nachfolgenden Prüfung und Entscheidungen in der ersten Prüfung in der Domäne „körperliche Untersuchung“ (siehe Abbildung 2, Feld 2c).

Abbildung 2 zeigt, dass ein signifikanter Zusammenhang zwischen den Entscheidungen in der ersten Prüfung und der OSCE-Punktzahl der nachfolgenden Prüfung besteht. Es sind einige Ausreißer enthalten (siehe Abbildung 2, Felder 2b, 2c, 3c, 4c); es besteht jedoch ein allgemeiner prädiktiver Zusammenhang. Entscheidungen für Bestehensnoten in der ersten Prüfung führten zu durchgängig

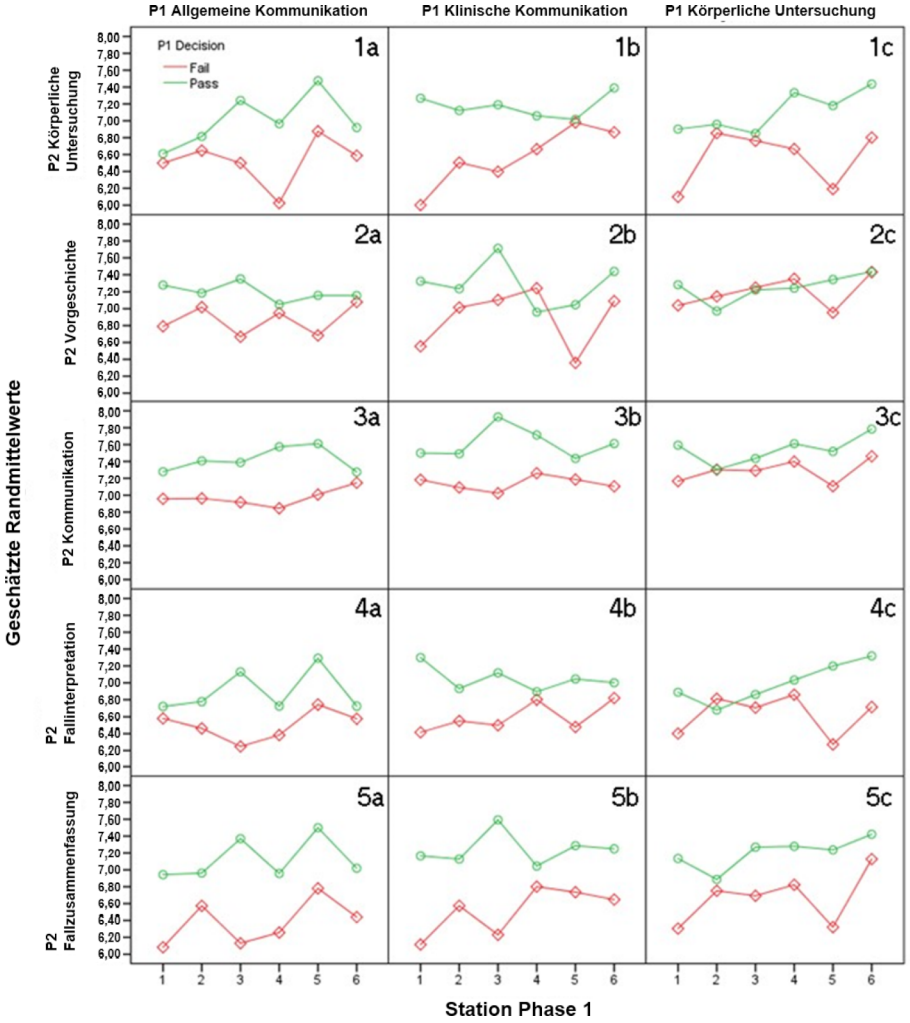

Abbildung 2: Der Vorhersagewert einer Entscheidung in der ersten Prüfung pro Bewertungsdomäne für in der nachfolgenden OSCE-Prüfung erhaltenen Noten pro Bewertungskriterium

besseren Noten als Entscheidungen für Nichtbestehensnoten.

Die ANOVA zeigt, dass dieses prädiktive Verhältnis mit den Entscheidungen in der ersten Prüfung speziell unabhängig von der Bewertungsstation in Zusammenhang steht. Die Ergebnisse deuten darauf hin, dass die Entscheidungen in der ersten Prüfung gerechtfertigt waren, da anhand vorheriger Noten die zukünftige Leistung vorhergesagt werden können sollte; dies ist basierend auf den Entscheidungen in der ersten Prüfung gelungen.

6. Diskussion

Entscheidungen in der ersten Prüfung weisen einen prädiktiven Zusammenhang auf, wenn sie für nachfolgende Prüfungen innerhalb einer Kohorte angewendet werden. Diese Vorhersagevalidität ist höher, wenn die Bewertungsdomänen der ersten Prüfung mit den Bewertungskriterien der nachfolgenden Prüfung verwandt sind, als wenn die Domänen/Kriterien weniger verwandt sind (siehe Tabelle 2; siehe Abbildung 1 und Abbildung 2).

Ein signifikanter Zusammenhang zwischen Entscheidungen in der ersten Prüfung und Noten in der nachfolgenden Prüfung besteht für die Entscheidungen in der ersten Prüfung in den Domänen „allgemeine Kommunikation“ und „klinische Kommunikation“ und die Noten in der nachfolgenden Prüfung in der Domäne „Vorgeschichte“ (siehe Tabelle 2; siehe Abbildung 1, Feld b; siehe Abbildung 2, Felder 2a–2b). Demgegenüber besteht für Entscheidungen in der ersten Prüfung in der Domäne „körperliche Untersuchung“ kein signifikanter Zusammenhang mit den Noten in der nachfolgenden Prüfung in der Domäne „Vorgeschichte“ (siehe Tabelle 2; siehe Abbildung 1, Feld b; siehe Abbildung 2, Feld 2c). Dies ergibt Sinn, da in diesen Domänen unterschiedliche, in den Domänen „Kommunikation“ und „Vorgeschichte“ jedoch ähnliche Fähigkeiten bewertet werden.

Auch wenn alle drei Bewertungsdomänen der ersten Prüfung in signifikantem Zusammenhang zu den in der nachfolgenden Prüfung in der Domäne „Kommunikation“ erhaltenen Noten steht, sind die Entscheidungen in der ersten Prüfung in den Domänen „allgemeine Kommunikation“ und „klinische Kommunikation“ wesentlich stärkere Prädiktoren als Entscheidungen in der ersten Prüfung in der Domäne „körperliche Untersuchung“ (Cohens d=0,725, 0,691 bzw. 0,388; siehe Tabelle 2; siehe Abbildung 1, Feld c; siehe Abbildung 2, Felder 3a–3c). Dies

zeigt, dass, auch wenn der prädiktive Zusammenhang bei den meisten Domänen besteht, er jedoch bei verwandten Domänen am stärksten ist.

Aufgrund der in den OSCE-Prüfungen der zweiten Phase erforderlichen Fallspezifität ist für die Fallinterpretation sowohl eine kompetente Leistung innerhalb einer Station, um die relevanten Informationen zu eruieren, als auch grundlegendes klinisches Wissen, um Fallergebnisse zu erhalten und sie intelligent zu interpretieren, notwendig. Dies zeigt die hohe Effektstärke in Zusammenhang mit den Noten in der nachfolgenden OSCE-Prüfung in den Domänen „Fallinterpretation“ und „Fallzusammenfassung“ (siehe Tabelle 2). Die medizinische Fakultät der UNSW hat festgelegt, dass eine gute Fallzusammenfassung auf mehreren in der OSCE-Prüfung der zweiten Phase bewerteten Faktoren beruht, einschließlich klarer/präziser allgemeiner Kommunikation, angemessener klinischer Fachsprache, Identifikation signifikanter Fallergebnisse und Aufzeigen von Differenzialdiagnosen [21].

Unveränderte Noten (Grenznoten) sind alle identisch und werden anhand der OBM2-basierten Entscheidung neu eingestuft. Ein prädiktiver Zusammenhang ist nur zu erwarten, wenn diese Entscheidungen valide sind. Wiederkehrende signifikante Zusammenhänge bei verschiedenen Bewertungsdomänen/-kriterien (siehe Abbildung 2) deuten darauf hin, dass diese Voraussagbarkeit kein zufälliges Ereignis ist. Zwischen den neueingestuften Noten und den zukünftigen Noten besteht ein prädiktiver Zusammenhang; derartige prädiktive Zusammenhänge sind auch in der Literatur zu finden [19]. Da die Entscheidungen diese Erwartungen widerspiegeln, vor allem bei verwandten Bewertungsdomänen/-kriterien und weniger bei weniger verwandten Bewertungsdomänen/-kriterien, erhöht sich die Validität der Entscheidungen.

Störgrößen wie der Prüfer, der Prüfungsstandort und die Stationen, an denen der Student geprüft wird, können Einfluss nehmen. Jede dieser Störgrößen wird im Folgenden besprochen.

An der medizinischen Fakultät der UNSW werden verschiedene Organisationsstrategien angewendet, um Urteilsverzerrungen zu minimieren und Urteilsfehler zu vermeiden. Für die an der UNSW durchgeführten OSCE-Prüfungen werden die Prüfer zufällig ausgewählt und den Prüfungsstandorten zugeteilt. Die Gutachter rotieren zwischen den verschiedenen Standorten, und es kommen externe Gutachter zum Einsatz [24]. Dadurch ist es höchst unwahrscheinlich, dass ein Student in beiden aufeinanderfolgenden Jahren vom selben Prüfer bewertet wird.

Daten der UNSW zeigen, dass es keinen signifikanten Unterschied bei der Leistung in den OSCE-Prüfungen zwischen den verschiedenen Prüfungsstandorten gibt [24]. Weiterhin werden die Studenten bei jeder OSCE-Prüfung den Prüfungsstandorten zufällig zugeteilt. Daher werden sie in den aufeinanderfolgenden Jahren nicht zwangsläufig am selben Standort geprüft.

Die OSCE-Prüfungen der ersten und zweiten Phase sind so konzipiert, dass sie verschiedenen Studienplänen gerecht werden und verschiedene Fähigkeiten bewertet werden [21]. An den OSCE-Stationen, an denen die Studenten bewertet werden, wird nicht dieselbe Fähigkeit oder dasselbe Wissen geprüft. Daher wird durch die Stationen, an denen die Studenten in der ersten OSCE-Prüfung bewertet werden, der Zusammenhang zwischen den Entscheidungen in der ersten Prüfung und den Noten in der nachfolgenden OSCE-Prüfung nicht verändert. Zusätzlich zeigen die Ergebnisse der ANOVA, dass bei keinen Bewertungsdomänen/-kriterien ein signifikanter Zusammenhang zwischen der Station in der ersten Phase und den Prüfungsnoten in der zweiten Phase besteht.

Nach Ausschluss dieser Variablen (Prüfer, Prüfungsstandort und Prüfungsstationen) ist es evident, dass der Großteil der prädiktiven Natur mit den Entscheidungen in Zusammenhang steht.

Dies stützt die Entscheidungen zur Neueinstufung von Grenznoten als eindeutige Bestehens- oder Nichtbestehensnoten. Die Validität der Entscheidungen wurde durch eine Reihe robuster statistischer Tests festgestellt. Dieser Bericht stützt zusammen mit früheren Studien die Validität dieser Entscheidungen [7], [14], [17]. Somit beseitigen diese Entscheidungen Unsicherheiten der Prüfer bei Grenzpunktzahlen. Dadurch kann die Objektivität bei der Neueinstufung von Grenznoten als eindeutige Bestehens- oder Nichtbestehensnoten erhöht werden.

Eine Einschränkung der Studie besteht darin, dass Daten nur einer Kohorte von Entscheidungen an einer Universität verwendet wurden. Die Bedeutung und die Reliabilität der Studie könnten verbessert werden, indem dieselben Tests für OSCE-Daten aufeinanderfolgender Jahre von verschiedenen Kohorten und an verschiedenen Universitäten durchgeführt würden, eine Wiederholung für diese Kohorte nach der dritten OSCE-Prüfung des Programms durchgeführt würde oder die OBM2 mit anderen Standard-Setting-Verfahren verglichen würde. All dies kann in zukünftigen Studien untersucht werden.

7. Schlussfolgerung

Es konnte bereits gezeigt werden, dass die Entscheidungen effizient, reliabel, belastbar und praktikabel sind. Weiterhin konnte in früheren Studien gezeigt werden, dass die Entscheidungen eine akzeptable Validität aufweisen. Die vorliegende Studie ist die erste Studie, die die Vorhersagevalidität der Entscheidungen zeigt und so die Validität der Entscheidungen zusätzlich stützt. Diese Ergebnisse können das Vertrauen der Prüfer bei folgenreichen Entscheidungen zur Neueinstufung von Grenznoten stärken.

In weiteren Untersuchungen können die bisher unbekannten Grenzen der OBM2 herausgestellt werden. Eine ähnliche Validierungsstudie kann durchgeführt werden, wenn die Daten der OSCE-Prüfung der dritten Phase für diese Kohorte verfügbar sind (im Jahr 2020), um zu untersuchen, ob die Vorhersagevalidität auch bei einer dritten nachfolgenden Prüfung ähnlich ist. Weiterhin kann die OBM2 innerhalb verschiedener Kontexte und für verschiedene Prüfungsformen getestet werden.

Interessenkonflikt

Die Autor*innen erklären, dass sie keinen Interessenkonflikt im Zusammenhang mit diesem Artikel haben.

Literatur

1. Rendel S, Foreman P, Freeman A. Licensing exams and judicial review: the closing of one door and opening of others? Br J Gen Pract. 2015;65(630):8-9. DOI: 10.3399/bjgp15X683029
2. Richard H, Sen GT, Jan V. The practical value of the standard error of measurement in borderline pass/fail decisions. Med Educ. 2008;42(8):810-815. DOI: 10.1111/j.1365-2923.2008.03103.x
3. Yudkowsky R, Tumuluru S, Casey P, Herlich N, Ledonne C. A Patient Safety Approach to Setting Pass/Fail Standards for Basic Procedural Skills Checklists. Simul Healthc. 2014;9(5):277-282. DOI: 10.1097/SIH.0000000000000044
4. Cizek GJ, Bunch MB. Standard setting: A guide to establishing and evaluating performance standards on tests. Thousand Oaks (CA): SAGE Publications Ltd; 2006.
5. Ben-David MF. AMEE Guide No. 18: Standard setting in student assessment. Med Teach. 2000;22(2):120-130. DOI: 10.1080/01421590078526
6. Phillips G. Technical Issues in Large-Scale Performance Assessment. Washington: U.S. Department of Education; 1996.
7. Shulruf B, Coombes L, Damodaran A, Freeman A, Jones P, Lieberman S, Poole P, Rhee J, Wilkinson T, Harris P. Cut-scores revisited: feasibility of a new method for group standard setting. BMC Med Educ. 2018;18(1):126. DOI: 10.1186/s12909-018-1238-7
8. Shulruf B, Wilkinson T, Weller J, Jones P, Poole P. Insights into the Angoff method: results from a simulation study. BMC Med Educ. 2016;16:134. DOI: 10.1186/s12909-016-0656-7
9. Hurtz GM, Hertz NR. How Many Raters Should be Used for Establishing Cutoff Scores with the Angoff Method? A Generalizability Theory Study. Educ Psychol Measurement. 1999;59(6):885-897. DOI: 10.1177/00131649921970233
10. Shulruf B, Turner R, Poole P, Wilkinson T. The Objective Borderline method (OBM): a probability-based model for setting up an objective pass/fail cut-off score for borderline grades in medical education programmes. Adv Health Sci Educ Theory Pract. 2013;18(2):231-144. DOI: 10.1007/s10459-012-9367-y
11. Wood T, Humphrey-Murto S, Norman G. Standard Setting in a Small Scale OSCE: A Comparison of the Modified Borderline-Group Method and the Borderline Regression Method. Adv Health Sci Educ Theory Pract. 2006;11(2):115-122. DOI: 10.1007/s10459-005-7853-1
12. Behuniak P, Archambault F, Gable R. Angoff and Nedelsky Standard Setting Procedures: Implications for the Validity of Proficiency Test Score Interpretation. Educ Psychol Measurement. 1982;42(1):247-255. DOI: 10.1177/0013164482421031
13. Poggio JP. An Empirical Investigation of the Angoff, Ebel and Nedelsky Standard Setting Methods. In: 65th Annual Meeting of the American Educational Research Association; 1981 Apr 13-17; Los Angeles, CA, United States. Zugänglich unter/available from: https://eric.ed.gov/?id=ED205552
14. Shulruf B, Poole P, Jones P, Wilkinson T. The Objective Borderline Method: a probabilistic method for standard setting. Ass Eval High Educ. 2015;40(3):420-438. DOI: 10.1080/02602938.2014.918088
15. Shulruf B, Adelstein BA, Damodaran A, Harris P, Kennedy S, O'Sullivan A, Taylor S. Borderline grades in high stakes clinical examinations: resolving examiner uncertainty. BMC Med Educ. 2018;18(1):272. DOI: 10.1186/s12909-018-1382-0
16. Shulruf B, Damodaran A, Jones P, Kennedy S, Mangos G, O'Sullivan A, Rhee J, Tayler S, Velan G, Harris P. Enhancing the defensibility of examiners' marks in high stake OSCEs. BMC Med Educ. 2018;18(1):10. DOI: 10.1186/s12909-017-1112-z
17. Shulruf B, Booth R, Baker H, Bagg W, Barrow M. Using the Objective Borderline Method (OBM) to support Board of Examiners' decisions in a medical programme. J Furth High Educ. 2017;41(3):425-434. DOI: 10.1080/0309877X.2015.1117603
18. Garson D. Validity and Reliabiity. North Carolina: Statistical Publishing Associates; 2016.
19. Poole P, Shulruf B, Rudland J, Wilkinson T. Comparison of UMAT scores and GPA in prediction of performance in medical school: a national study. Med Educ. 2012;46(2):163-171. DOI: 10.1111/j.1365-2923.2011.04078.x
20. University of New South Wales, Faculty of Medicine. Phase 1 / Graduate Entry Clinical Skills Student Guide 2018. Kensington: The University of New South Wales; 2018.
21. University of New South Wales, Faculty of Medicine. Phase 2 Clinical SKills Guide 2018. Kensington: The University of New South Wales; 2018.
22. IBM Corporation. IBM SPSS Statistics for Windows. 24 ed. Armonk, NY: IBM Corporation; 2016.
23. Wilson D. Practical Meta-Analysis Effect Size Calculator. Fairfax: George Mason University; 2018.
24. Medical School Accreditation Committee. Accreditation of University of New South Wales Faculty of Medicine. Kingston: Australia Medical Council Limited; 2018.

Korrespondenzadresse:
Boaz Shulruf
University of New South Wales, Office of Medical Education, Sydney, Australien
b.shulruf@unsw.edu.au

Bitte zitieren als
Klein Nulend R, Harris P, Shulruf B. Predictive validity of a tool to resolve borderline grades in OSCEs. GMS J Med Educ. 2020;37(3):Doc31. DOI: 10.3205/zma001324, URN: urn:nbn:de:0183-zma0013243

Artikel online frei zugänglich unter
https://www.egms.de/en/journals/zma/2020-37/zma001324.shtml

Eingereicht: *18.03.2019*
Überarbeitet: *19.11.2019*
Angenommen: *07.01.2020*
Veröffentlicht: *15.04.2020*