Overcome the 60% passing score and improve the quality of assessment

Abstract

It is not unusual for institutions around the world to have fixed standards (e.g., 60%) for all of their examinations. This creates problems in the creation of examinations, since all of the content has to be chosen with an eye toward this fixed standard. As a result, the validity of the decisions based on these examinations can be adversely influenced, making them less useful for their intended purposes.

Over the past several decades, many institutions have addressed this problem by using standard setting methods which are defensible, acceptable, and credible [1], [2]. Many methods are available and the major reasons to use them is to ensure that test content is appropriately selected and to be as fair to the students and other test users as possible [2], [3].

One barrier to the wider use of these methods is that some institutions object to the fact that the fixed standard (e.g., 60%) has not been applied. However, it is possible to rescale the passing score so that it is equal to the fixed standard, and then apply that same rescaling calculation to all of the test scores. This ensures that the institutional guidelines are not violated and allows the application of accepted methods of standard-setting. In turn, the application of these methods allow the content of the test to be selected without regard to a fixed standard, increases the validity of the decisions being made, and ensures a fairer and more accurate test of students.

Ara Tekian[1]
John Norcini[2]

1 University of Illinois, College of Medicine at Chicago, Chicago, USA

2 Foundation for Advancement of International Medical Education and Research , Philadelphia, USA

Commentary

Over the past several decades, many authors have advocated for setting standards for passing a test which are defensible, acceptable, and credible [1], [2]. Many methods are available and the major reasons to use them is to ensure that test content is appropriately selected and to be as fair to the students as possible [2], [3].

The implementation of a standard setting method moves the passing standard away from a fixed number (i.e., 60%) to a cut score which can vary depending on the difficulty of the test. For example, when two forms of an exam are administered, where Form A is slightly more difficult than Form B, setting the same passing standard for both will give an unfair advantage to students taking Form B – this threatens the validity of the test, by passing candidates who may not be qualified simply because of the test characteristics (difficulty) of the exam. Setting a relative passing standard (e.g., mean – 2 x standard deviation) does not overcome this problem, because the ability of students can change from year to year. Application of an absolute standard-setting method that relies on the judgments of a panel of expert is preferred.

Examinations may be administered for formative and summative purposes. Formative assessment is focused on providing feedback to the students. However, summative examinations which are focused on making decisions about students' competence might have a significant impact on a student's career pathway.

In all of the health professions, including medicine, the undergraduate students need to be judged on their mastery of their professional content. They need to be competent in terms of knowledge and performance, and therefore, need to be assessed against a set of criteria or standards. The standard setting processes utilized for such purposes generate absolute standards, in contrast to relative standards where students are judged against each other. These standards are considered absolute, because they are expressed in terms of how much content the students need to know and thus theoretically all could pass or fail. Therefore, the success rate for any examination might vary depending on the passing score established by content experts.

There are various methods of setting absolute standards, and the judgments might be focused on either the items, know as item-based (Angoff, Ebel [4], [5]), or on the examinee, that is examinee-based (Borderline or Contrasting group methods [6]). There are also compromise methods of setting standards, for example the Hofstee method, where judgments about how much needs to be known are combined with the relative performance of the students, also popularly known as the relative-absolute compromise method [7]. The passing scores might vary depending on the choice of method [2], [8].

In some methods, the judgments underlying standard setting are based on the definition of a hypothetical borderline student who would have a 50%-50% probability of passing the exam. The description of the borderline student is based on a consensus definition in the content areas represented by the blueprint and generated by the expert panel. This definition is based on a hypothetical candidate who on a given day would pass the exam and on a different day, fail the exam; the competency demonstrated by this candidate should represent uncertainty about the qualifications and attributes required for a passing candidate; it could also include descriptions of "forgivable" qualities that the candidate may not yet have, but over the course of his or her training, continue to master.

Although the Angoff method for standard setting is a very popular method used in many medical schools, attention must be paid to the selection of judges, since their level of expertise, and the ability to answer the exam items correctly may affect the passing score [9]. Choice of credible judges and their calibration are equally important in standard setting, particularly during discussions of the borderline-examinee, and during training sessions/ exercises where the probability of a borderline-examinee answering or performing a checklist item correctly is being estimated [2].

Some institutions have started using the Angoff or other methods to set the passing score for OSCE examinations. They have adopted this strategy for OSCE examinations because the institution will not notice or object that the 60% passing score is not applied, or they have manipulated the difficulty of the OSCE examination in such a way as to have a passing score of 60% without any adjustments. However, the results of a standard setting exercise can also be rescaled to the passing score determined by the university, such as 60%. This can be achieved, without violating the institutional guidelines, by converting the raw passing score of an exam, determined through standard setting into a rescaled passing standard of 60% [10]. The scores of students are then rescaled as well. For example, if the passing score using the Angoff method for an MCQ examination was calculated to be 54%, that score is rescaled to 60% and then the examination is rescaled as well, so that the students and passing score are consistent with the '60%' policy. Therefore, following a standard setting method will overcome the capricious nature of assigning a passing score of 60%, by simply transposing the true passing score to the institutionally required cut-off score.

There is no "gold standard" for setting passing scores [11], [2]. Usually the choice of the standard setting method is based on the available resources and the practical realities of the educational environment. As such, it is critical to document all procedures used in establishing the passing standard, especially in a language the school is willing to make public. Every effort should be exerted to use one of the standard setting approaches described in the literature so as to make the passing scores defensible and meaningful. To accomplish this, there needs to be intensive faculty development, so that everyone understands the importance and consequences of setting standards, and are trained and calibrated to set passing scores. Moreover, faculty selected to participate in the standard setting process need to be representative and acceptable to the stakeholders. This may also require institutional change and one way to accomplish this is to hold meaningful discussions with institutional leaders about the rationale for standard-setting and evidence behind it. Using standard setting procedures necessitates changing the institutional assessment culture and promoting fairness and justice in measuring the competence of the students.

Competing interests

The authors declare that they have no competing interests.

References

1. Norcini JJ. Setting standards on educational tests. Med Educ. 2003;37(5):464-946. DOI: 10.1046/j.1365-2923.2003.01495.x
2. Downing SM, Tekian A, Yudkowsky R. Procedures for establishing defensible absolute passing scores on performance examinations in health professions education. Teach Learn Med. 2006;18(1):50-57. DOI: 10.1207/s15328015tlm1801_11
3. Yudkowsky R, Downing SM, Tekian A. Standard setting. In: Downing SM, Yudkowsky R (Hrsg). Assessment in health professions education. New York/London: Routledge; 2009. S.119-148.
4. Angoff WH. Scales, norms, and equivalent scores. In: Thorndike RL (Hrsg). Educational measurement. 2nd ed. Washington, DC: American Council on Education; 1971. S.508-600.
5. Ebel RL. Essentials of educational measurement. 2nd ed. Englewood Cliffs, NJ: Prentice Hall; 1972.
6. Livingston SA, Zieky MJ. Passing scores: A manual for setting standards of performance on educational and occupational tests. Princeton, NJ: Educational Testing Service; 1982.
7. Hofstee WK. The case for compromise in educational selection and grading. In: Anderson SB, Helmick JS (Hrsg). On educational testing. San Francisco: Jossey-Bass, 1983. S.107-127.
8. Bouriscot K, Roberts T, Pell G. Standard setting for clinical competence at graduation from medical school: A comparison of passing scores across five medical schools. Adv Health Sci Educ Theory Pract. 2006;11(2):173-183. DOI: 10.1007/s10459-005-5291-8
9. Verheggen MM, Muijtjens AM, Van Os J, Schuwirth LW. Is an Angoff standard an indication of minimal competence of examinees or of judges? Adv Health Sci Educ Theory Pract. 2008;13(2):203-211. DOI: 10.1007/s10459-006-9035-1
10. Kolen MJ, Brennan R L. Test equating, scaling, and linking: Methods and practices. 2nd Edition. New York: Springer-Verlag; 2004. DOI: 10.1007/978-1-4757-4310-4
11. Friedman M. AMEE Guide No. 18: Standard setting in student assessment. Med Teach. 2000;22(2):120-130. DOI: 10.1080/01421590078526

Corresponding authors:
Ara Tekian, PhD, MHPE
University of Illinois, College of Medicine at Chicago, 808 South Wood Street (MC-783) , Chicago, IL 60612-7302, USA
tekian@uic.edu
Dr. John Norcini, Ph.D
Foundation for Advancement of International Medical Education and Research, 3624 Market Street
Philadelphia, PA 19104-2685 USA
JNorcini@Faimer.org

Please cite as
Tekian A, Norcini J. Overcome the 60% passing score and improve the quality of assessment. GMS Z Med Ausbild. 2015;32(4):Doc43. DOI: 10.3205/zma000985, URN: urn:nbn:de:0183-zma0009859

This article is freely available from
http://www.egms.de/en/journals/zma/2015-32/zma000985.shtml

Received: *2014-01-14*
Revised: *2014-01-14*
Accepted: *2014-05-26*
Published: *2015-10-15*

OPEN ACCESS

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

Die 60% Bestehensgrenze überwinden und die Qualität der Prüfungen verbessern

Zusammenfassung

Es ist weltweit häufig üblich, formale Bestehensgrenzen (z. B. 60%) für alle Prüfungen festzulegen. Dies führt zu Problemen bei der Entwicklung von Prüfungen, da der gesamte Inhalt der Prüfung im Hinblick auf diesen festgelegten Standard abzustimmen ist. Infolgedessen kann die Validität der Entscheidungen, die auf diesen Prüfungen fußen, nachteilig beeinflusst werden und den Nutzen für ihren beabsichtigten Verwendungszweck einschränken.

Im Laufe der letzten Jahrzehnte begegneten viele Institute diesem Problem durch die Verwendung von Verfahren des „Standard-Settings“, also Festlegungen der Bestehensgrenzen, die vertretbar, akzeptabel und zuverlässig sind [1], [2]. Hierzu steht eine Vielzahl an Methoden zur Verfügung und die Hauptgründe, diese zu nutzen, bestehen darin, sicherzustellen, dass der Testinhalt adäquat ausgewählt und so fair wie möglich gegenüber den Studierenden und anderen Testanwendern ist [2], [3].

Ein Hindernis für die breitere Anwendung dieser Verfahren ist, dass vielerorts die Tatsache beanstandet wird, dass dabei die formal festgesetzte Bestehensgrenze (z. B. 60%) nicht eingehalten wird. Es ist jedoch möglich, durch Umskalierungen die Bestehensgrenze neu zu definieren, so dass diese dem festgelegten Standard entspricht und dieselbe Skalierung auf alle Testergebnisse anwendet. Dies stellt sicher, dass die institutionellen Richtlinien nicht verletzt werden und ermöglicht die Anwendung der anerkannten Verfahren des Standard-Settings. Im Gegenzug gestattet die Anwendung dieser Verfahren die Auswahl der Prüfungsinhalte ohne Berücksichtigung des festgelegten Standards, was die Validität der getroffenen Entscheidungen erhöht und eine gerechtere und genauere Prüfung der Studierenden ermöglich.

Ara Tekian[1]
John Norcini[2]

1 University of Illinois, College of Medicine at Chicago, Chicago, USA

2 Foundation for Advancement of International Medical Education and Research , Philadelphia, USA

Kommentar

Im Laufe der letzten Jahrzehnte wurde von vielen Autoren vorgeschlagen, Standards für das Bestehen einer Prüfung zu setzen, die vertretbar, akzeptabel und zuverlässig sind [1], [2]. Hierzu stehen viele Verfahren zur Verfügung und die Hauptgründe, diese anzuwenden, ist es, sicher zu stellen, dass die Prüfungsinhalte adäquat ausgewählt werden und so fair wie möglich gegenüber den Studierenden sind [2], [3].

Die Implementierung eines Standard-Setting-Verfahrens bewegt die Bestehensgrenze weg von einer festgelegten Zahl (z. B. 60%) hin zu variablen Notengrenzen, die von der Testschwierigkeit abhängig sind. Bespiel: Wenn zwei Varianten einer Prüfung verwendet werden, von welcher Prüfung A geringfügig schwieriger ist als Prüfung B, ergibt sich daraus, sofern dieselbe Bestehensgrenze für beide Prüfungen angewendet wird, ein unfairer Vorteil für Studierende mit Prüfung B. Dies gefährdet die Validität der Prüfung, da nicht qualifizierten Kandidaten aufgrund der Testcharakteristika (Schwierigkeit) der Prüfung bestehen. Das Setzen einer relativen Bestehensgrenze (z. B. Mittelwert – 2 Standardabweichungen) löst dieses Problem nicht, da sich das Können der Studierenden von Jahr zu Jahr ändern kann. Die Anwendung eines absoluten Standard-Setting-Verfahrens, das auf dem Urteil eines Expertengremiums beruht, ist zu bevorzugen.

Prüfungen können für formative und summative Zwecke eingesetzt werden. Formative Prüfungen konzentrieren sich darauf, den Studierenden Feedback zu geben. Summative Prüfungen, die auf Entscheidungen über das Vorhandensein erforderlicher Kompetenzen fokussiert sind, können einen entscheidenden Einfluss auf den Karriereweg von Studierenden haben.

In allen Gesundheitsberufen, die Medizin eingeschlossen, müssen die Studierenden auf der Grundlage ihres beruflichen Könnens und Wissens beurteilt werden. Sie müssen in Bezug auf Wissen und Leistung kompetent sein und deshalb anhand einer Menge von Kriterien oder Standards geprüft werden. Das für diesen Zweck genutzte Standard-Setting-Verfahren generiert absolute Standards im Gegensatz zu relativen (normorientierten) Standards, die Studierende miteinander vergleichen. Diese Standards werden als „absolut“ bezeichnet, weil sie als Bedingungen formuliert werden, welche Inhalte die Studierenden wis-

sen müssen und in der Theorie folglich jeder „bestehen“ oder „nicht bestehen“ kann. Daher kann die Erfolgsrate für jede Prüfung variieren, abhängig von der Bestehensgrenze, die von einem Expertengremium festgesetzt wurde.

Es gibt eine Vielzahl von Verfahren zur Festsetzung absoluter Standards, diese können auf die Testaufgaben (testzentriert: Angoff, Ebel [4], [5]) oder auf den Prüfling (prüflingszentriert: Borderline-Methode oder Kontrast-Gruppen-Methode) fokussiert sein [6]. Es existieren auch Standard-Setting-Verfahren, die einen Kompromiss zwischen diesen darstellen, z. B. die Hofstee-Methode („relative-absolute compromise method“). Dabei werden Festlegungen, welche Inhalte gewusst werden müssen mit der relativen Leistung der Studierenden kombiniert. [7]. Abhängig vom gewählten Verfahren können die Bestehensgrenzen variieren [2], [8].

Bei einigen Verfahren basieren die Beurteilungen der zugrunde liegenden Standard-Settings auf der Definition eines hypothetischen Borderline-Prüfungskandidaten, der mit einer 50%-igen Wahrscheinlichkeit die Prüfung besteht. Die Beschreibung des Borderline-Prüfungskandidaten basiert auf einer Konsensentscheidung bezüglich der Inhaltsbereiche, die durch einen Blueprint repräsentiert und von einem Expertengremium getroffen wurden. Diese Definition basiert auf einem hypothetischen Prüfungskandidaten, der an einem bestimmten Tag die Prüfung bestehen würde und sie an einem anderen Tag nicht bestehen würde; die Kompetenz, die von einem solchen Prüfungskandidaten gezeigt wird, soll die Unsicherheit über die zum Bestehen erforderlichen Qualifikationen und Eigenschaften repräsentieren; sie können auch Beschreibungen von „verzeihlichen“ Qualitäten/Eigenschaften umfassen, die der Kandidat zwar noch nicht hat, aber im Laufe seiner Ausbildung noch erlangen kann.

Obwohl die Angoff-Methode als Standard-Setting an vielen Medizinischen Fakultäten eine sehr beliebte Methode ist, muss die Aufmerksamkeit auf die Auswahl der Experten gerichtet werden, da ihre Expertise und ihre Fähigkeit, die Prüfungsfragen korrekt zu beantworten, die Bestehensgrenze beeinflussen kann [9]. Die Wahl zuverlässiger Experten und ihrer Kalibrierung sind gleichermaßen für das Standard-Setting wichtig. Dies gilt besonders für die Diskussion des Borderline-Prüfungskandidaten und im Verlauf von Probedurchgängen, in denen die Wahrscheinlichkeit eines Borderline-Kandidaten, ein Checklistenitem zu beantworten oder korrekt auszuführen, abgeschätzt wird [2].

An einigen Institutionen wurde begonnen, die Angoff-Methode oder andere Verfahren zu verwenden, um die Bestehensgrenze für OSCE-Prüfungen festzulegen. Sie haben diese Strategie für OSCE-Prüfungen gewählt, weil nicht bemerkt oder nichts dagegen eingewandt wurde, von der 60%-Bestehensgrenze abzuweichen, oder aber es wurde die Schwierigkeit der OSCE-Prüfung so beeinflusst, dass eine Bestehensgrenze von 60% ohne Anpassungen verwendet werden kann. Es kann jedoch auch das Ergebnis eines Standard-Settings auf die durch die Universität festgelegte Bestehensgrenze, wie etwa 60%, reskaliert werden: Dies kann ohne Verletzung der Institutsrichtlinien durch die Umwandlung der Rohwert-Bestehensgrenze einer Prüfung, die durch ein Standard-Setting festgelegt wurde, in eine neuskalierte Bestehensgrenze von 60% erreicht werden [10]. Die erreichten Punktzahlen der Studierenden werden gleichermaßen neu skaliert. Beispiel: Wurde die Bestehensgrenze für eine MCQ-Prüfung auf 54% mittels der Angoff-Methode ermittelt, wird dieser Wert der Bestehensgrenze auf 60% neuskaliert. Ebenso wird die Prüfung neuskaliert, so dass die Punkte der Studierenden und die Bestehensgrenze mit der ‚60%‘-Regel konsistent sind. Damit kann nach einem Standard-Setting-Verfahrens die willkürliche Natur einer festen Bestehensgrenze von 60% durch eine einfache Neuskalierung der ursprünglichen Bestehensgrenze an die institutionell geforderte überwunden werden.

Für die Festlegung der Bestehensgrenze gibt es keinen „Goldstandard“ [11], [2]. Üblicherweise basiert die Wahl des Standard-Setting-Verfahrens auf den verfügbaren Ressourcen und der Ausstattung für die Ausbildung. Es ist entscheidend, alle angewandten Verfahren zur Festlegung der Bestehensgrenze zu dokumentieren und dies in der für die Fakultät üblichen Form publik zu machen. Es sollte jede Anstrengung unternommen werden, um eines, der in der Literatur beschrieben Standard-Setting-Verfahren einzusetzen, um so die Bestehensgrenze vertretbar und bedeutsam zu machen. Um dies zu erreichen, ist eine intensive Fakultätsentwicklung erforderlich, damit jeder die Wichtigkeit und die Konsequenzen der Festsetzung von Standards versteht, trainiert und befähigt ist, Bestehensgrenzen festzulegen. Darüber hinaus müssen die Dozenten, die am Prozess des Standard-Settings teilnehmen, für die Beteiligten repräsentativ sein und von ihnen akzeptiert werden. Dies könnte auch einen institutionellen Wandel verlangen. Ein Weg, dies zu erreichen, ist es Diskussionen mit der Fakultätsleitung über das Rational des Standard-Settings und die dahinterstehende Evidenz zu führen. Die Verwendung von Standard-Setting-Verfahren verlangt eine Änderung der Prüfungskultur der Institutionen sowie die Forderung nach Fairness und Gerechtigkeit bei der Kompetenzmessung der Studierenden.

Interessenkonflikt

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

Literatur

1. Norcini JJ. Setting standards on educational tests. Med Educ. 2003;37(5):464-946. DOI: 10.1046/j.1365-2923.2003.01495.x
2. Downing SM, Tekian A, Yudkowsky R. Procedures for establishing defensible absolute passing scores on performance examinations in health professions education. Teach Learn Med. 2006;18(1):50-57. DOI: 10.1207/s15328015tlm1801_11

3. Yudkowsky R, Downing SM, Tekian A. Standard setting. In: Downing SM, Yudkowsky R (Hrsg). Assessment in health professions education. New York/London: Routledge; 2009. S.119-148.

4. Angoff WH. Scales, norms, and equivalent scores. In: Thorndike RL (Hrsg). Educational measurement. 2nd ed. Washington, DC: American Council on Education; 1971. S.508-600.

5. Ebel RL. Essentials of educational measurement. 2nd ed. Englewood Cliffs, NJ: Prentice Hall; 1972.

6. Livingston SA, Zieky MJ. Passing scores: A manual for setting standards of performance on educational and occupational tests. Princeton, NJ: Educational Testing Service; 1982.

7. Hofstee WK. The case for compromise in educational selection and grading. In: Anderson SB, Helmick JS (Hrsg). On educational testing. San Francisco: Jossey-Bass, 1983. S.107-127.

8. Bouriscot K, Roberts T, Pell G. Standard setting for clinical competence at graduation from medical school: A comparison of passing scores across five medical schools. Adv Health Sci Educ Theory Pract. 2006;11(2):173-183. DOI: 10.1007/s10459-005-5291-8

9. Verheggen MM, Muijtjens AM, Van Os J, Schuwirth LW. Is an Angoff standard an indication of minimal competence of examinees or of judges? Adv Health Sci Educ Theory Pract. 2008;13(2):203-211. DOI: 10.1007/s10459-006-9035-1

10. Kolen MJ, Brennan R L. Test equating, scaling, and linking: Methods and practices. 2nd Edition. New York: Springer-Verlag; 2004. DOI: 10.1007/978-1-4757-4310-4

11. Friedman M. AMEE Guide No. 18: Standard setting in student assessment. Med Teach. 2000;22(2):120-130. DOI: 10.1080/01421590078526

Korrespondenzadressen:
Ara Tekian, PhD, MHPE
University of Illinois, College of Medicine at Chicago, 808 South Wood Street (MC-783) , Chicago, IL 60612-7302, USA
tekian@uic.edu
Dr. John Norcini, Ph.D
Foundation for Advancement of International Medical Education and Research, 3624 Market Street
Philadelphia, PA 19104-2685 USA
JNorcini@Faimer.org

Bitte zitieren als
Tekian A, Norcini J. Overcome the 60% passing score and improve the quality of assessment. GMS Z Med Ausbild. 2015;32(4):Doc43. DOI: 10.3205/zma000985, URN: urn:nbn:de:0183-zma0009859

Artikel online frei zugänglich unter
http://www.egms.de/en/journals/zma/2015-32/zma000985.shtml

Eingereicht: *14.01.2014*
Überarbeitet: *14.01.2014*
Angenommen: *26.05.2014*
Veröffentlicht: *15.10.2015*