establecido en nuestro país México y por lo tanto las Instituciones como Pemex, Secretaría de Comunicaciones y Transportes, Secretaría de la Defensa Nacional, Poder Judicial de la Federación, Instituciones Educativas, así como Clínicas y Hospitales debidamente acreditados solicitan al Médico General estar Certificado por CONAMEGE
El Consejo Nacional de Certificación en Medicina General A.C. conformado por los Consejos Estatales, tiene una estructura federada y cada Consejo cuenta con una Mesa Directiva, Presidente, Vicepresidente, Secretario, Tesorero, que tienen un periodo lectivo de tres años. Entre 15 a 25 Consejeros y un Cuerpo Consultivo formado por los Ex presidentes que con su experiencia apoyan las estrategias de difusión y realización de los Exámenes que tienen el mayor protocolo de ética y se realizan en todo el país en el mes de Febrero, Junio y Octubre de cada año. Con la posibilidad de realizar fechas extemporáneas cuando así se requiere.
Es una actividad Académica de gran responsabilidad en donde no hay conflicto de intereses ni fines de lucro.

BIBLIOGRAFIA

1.Cervantes Carreño M. Consejo Nacional de Certificación en Medicina General. Lecture presented at; 2017; Academia Nacional de Medicina de México.
2.Órgano Informativo del Comité Normativo Nacional de Medicina General. CONAMEGE: Presentación. Boletín del Médico General 2013;1(1):1.
3.Órgano Informativo del Comité Normativo Nacional de Medicina General. La Asociación Mexicana de Facultades y Escuelas de Medicina. Boletín del Médico General 2007;4(2):1.
4.Órgano Informativo del Comité Normativo Nacional de Medicina General. Un poco de historia del CONAMEGE. Boletín del Médico General 2013;1(1):2.
5.Órgano Informativo del Comité Normativo Nacional de Medicina General. CONAMEGE: Ayer y Hoy. Boletín del Médico General 2007;4(3):2.
6.Órgano Informativo del Comité Normativo Nacional de Medicina General. La certificación del Médico General. Boletín del Médico General 2005;2(3):1.
7.Órgano Informativo del Comité Normativo Nacional de Medicina General. Certificación. Boletín del Médico General 2006;3(2):2.
8.Órgano Informativo del Comité Normativo Nacional de Medicina General. El Examen de Certificación para el Médico General. Boletín del Médico General 2007;4(2):2.
9.Órgano Informativo del Comité Normativo Nacional de Medicina General. Requisitos para la Certificación en Medicina General. Boletín del Médico General 2009;6(1):4.
10.Órgano Informativo del Comité Normativo Nacional de Medicina General. La Profesionalización de Examen y la Búsqueda de la Idoneidad. Boletín del Médico General 2010;7(1):1.
11.Instituto de Ingeniería y evaluación Avanzada I.E.I.A órgano evaluador externo.

Artículo original

Standards of an Examination (cut points), Construction of Scenarios, Measurement of Impacts in the Decision Making Process

Raja G. Subhiyah

National Board of Medical Examiners

DIRECCIÓN PARA CORRESPONDENCIA: Raja G. Subhiyah, RSubhiyah@nbme.org.

RESUMEN

La Junta Nacional de Examinadores Médicos (NBME®) aplica estándares rigurosos para determinar el impacto en los candidatos que toman el Examen de Licencia Médica de los Estados Unidos (USMLE®) con el propósito de otorgar licencias a médicos en los EE. UU. Estos estándares se aplican a todos los niveles de desarrollo, administración y calificación de los exámenes. Los estándares se aplican a los siguientes procesos:
• Validez de las inferencias de los puntajes: contenido: el contenido probado debe ser apropiado, haciendo las preguntas correctas, formato del elemento, o diseño de prueba y diseño, o proceso por el cual se desarrolla la prueba, documentación
• Precisión de los puntajes: confiabilidad, focalización, información en el puntaje de corte, errores estándar
• Determinación y aplicación de los puntos de corte: métodos, el procedimiento Angoff modificado, error de clasificación errónea.
El enfoque principal está en el último estándar, aunque también se discutirán brevemente los dos primeros estándares. Se discutirán diferentes métodos para establecer un estándar de aprobación y se describirá el método utilizado para USMLE. También se presentarán errores de clasificación errónea y cómo minimizarlos.

Palabras clave: Configuración estándar, licencia, certificación, estándares.

ABSTRACT

The National Board of Medical Examiners (NBME®) applies rigorous standards for determining impact on candidates taking the United States Medical Licensure Examination (USMLE®) for the purpose of licensing medical doctors in the USA. These standards apply to all levels of developing, administering and scoring the examinations. The standards apply to the following processes which will be discussed:

• Validity of the inferences from the scores: content: the content tested must be appropriate, asking the right questions, item format , blueprinting and test design, process by which the test is developed, documentation.
• Precision of the scores: reliability, targeting, information at the cut score, standard errors.
• Determination and application of the cut-points: methods, the modified Angoff procedure, misclassification error.

The main focus will be on the last standard, although the first two standards will also be briefly discussed. Different methods of setting a passing standard will be discussed and the method used for USMLE described. Misclassification errors and how to minimize them will also be presented.

Key words: Standard Setting, Licensure, Certification, Standards.

INTRODUCCION

A standard usually depicts a minimum requirement. This requirement can represent a level, such as the level of complexity of a performed task, or it may represent a quantity, such as the amount of content mastered by the candidate.[1,8] In the medical certification and licensure fields, the most commonly occurring standards are expressed as minimum passing scores (MPS) on one or more examinations. Thus, a passing score implies that the candidate is proficient enough to meet the requirements for certification/licensure while a failing score implies that adequate proficiency is lacking. This scheme seems to apply to both knowledge and performance tests.[1,14]

In order to make appropriate and fair inferences from a score, it is necessary for that score itself to meet certain standards of quality. There are two main attributes that make a test score adequate for the purpose certification and licensure. First, the score must be meaningful and relevant. That is, the score must represent a measure of a quality that is relevant and essential to the substance of the certification or license, and therefore, will enable the user to make legitimate and appropriate inferences about the candidate based on the score. This attribute is briefly known as validity.[1,7]

Second, the score must measure the desired quality with sufficient precision. If the attribute we are measuring is not sufficiently precise, we cannot interpret the level of the score to have a useful meaning. This attribute of the score is usually depicted by the standard error of measurement (SEM).[1,7]

These two properties of a score are key elements that determine the quality of decisions made about candidates for certification and licensure. These decisions which have a profound impact on the careers and lives of the candidates should be treated with careful scrutiny that does justice to the importance that they warrant.

One more factor in determining minimum passing scores (MPS) begs attention: methodology. The method by which a standard is set may also impact the quality and nature of the decisions based on that standard. In general, there are two types of standards. Norm-based methods set the MPS relative to the performance of a certain group, usually known as the reference group. Criterion-based methods set the MPS to require a certain level or amount of a criterion regardless of actual performance of the reference group.

The rest of this document will discuss in more detail three standards for setting certification/licensure standards: validity, precision, and methods.

Standard I: Validity of Inference

While, in general, standards employed in medical certification and licensure programs in the United States have various requirements, a common and essential feature of most of these programs is the requirement to pass an examination.[2,4,5] It is on this feature that this paper focuses. Most of these required examinations measure knowledge about a content that is relevant and essential to the practice at hand. Thus, the main facet of validity that concerns these programs is content validity.

Hence, the question to be addressed in these cases: is the content measured by the examination relevant to safe and effective practice? Since the standard is expressed as a minimum passing score (MPS), it is one value on a score scale. We are assuming here that the scores quantify the attribute, in this case, medical knowledge, that is being measured. Thus, we believe that higher scores represent a larger degree of knowledge than lower scores.[7,19]

How do we establish the all-important relationship between scores and degree of knowledge? The key here is the process by which the examination is developed and the documentation of this process. As one might surmise, developing a licensing or certifying examination is not a simple matter, and requires several stages. The first stage in developing such an examination is arriving at a clear definition of the purpose of the examination and the inferences desired from the scores.[3,6] Why are we giving this examination, and what do we want to know from the scores?

Once the purpose and desired inferences are established, the content of the examination must be defined and organized. This may be achieved by a job analysis study or by a rapid blueprinting study. These studies are used to determine the knowledge profile that a practitioner needs to have in order to successfully practice. In short, they determine what content must be on the examination and in what proportions.[2]

The end result of either method is a blueprint (design) of the content of the examination. The blueprint is usually organized to have three essential components:

- Major content areas that break down the overall content domain into manageable cohesive areas of knowledge (e.g., in a Clinical Science exam, major areas can be: nutrition and digestive, endocrine and metabolic, skin & musculoskeletal),
- within each major area, the specific concepts that must be tested. Roughly, each specific element would be appropriate for one or two questions on the test, and
- weights to balance the relative importance of each major area in relation to the others. These weights may be expressed as the number of questions from each area. The weights can be refined to highlight certain specific areas that are especially important.

After the blueprint is developed, it can be used as a basis for giving item writers their assignments. Item writers must be content experts/practitioners of high quality and a good knowledge of item-authoring skills. It is common practice in the best examination programs to give all item writers a workshop in item-authoring skills. These skills eliminate flaws in the items and homogenize their format so that the test-taker is not confused. As a result, the examination captures the intended knowledge without distracting and irrelevant factors.

After receiving the items from the authors, they are edited for flaws, clarity, and format. Revised items are returned to the authors for review to ascertain that content has not been changed. After a few iterations, the items are pronounced acceptable, and placed in the pretesting pool and administered to candidates mixed with already established items. If the statistical performance of the items is acceptable, they will be placed in the live item pool from which the scored examination is compiled.

The criteria and conditions set in the blueprint are applied and an examination is created. The best programs have a committee of content experts review each form of the examination and make necessary substitutions. The examination is now ready to be administered. The administration of the examination should subject all candidates to equal conditions and time constraints. Irregular behavior such as copying must be prevented so that the examination accurately captures the knowledge of the candidates and scores reflect the intended proficiencies.[12]

Once the examination is administered, all candidate records are scored in exactly the same fashion. Scores are then computed and reported to candidates with any decisions that are based on them. The whole process of developing the examination attests to the validity of inferences derived from the scores.

Standard II: Precision

Precision of the scores will affect the ability to pin-point the true ability of the candidate. The purpose of the examination is to estimate the level of proficiency that a candidate has in this content. The content domain, which comprises all that a candidate must know in order to be certified is very large, and can be considered practically infinite. To make the estimate possible, we must sample the domain according to a well thought-out design (blueprint), and test the candidate on that sample (the examination). We assume that the examination consists of a balanced sample of the infinite content domain.[7,9]

If for example, a candidate correctly responds to 60% of the items on the examination, we assume that she or he knows approximately 60% of the content. The generalization from examination to content domain is not perfect, but includes a certain amount of uncertainty usually called measurement error. This error in the measurement depends on several factors of which the most important is sample size. The sample size in this case is the number of items on the examination.

Because of the error in measurement, if the same person takes parallel forms of the same examination multiple times, the obtained scores will vary from time to time. And if the number of tests is large, that person's scores will distribute normally around a mean, and the distribution will have a standard deviation. This standard deviation is called the Standard Error of Measurement (SEM). The SEM is important in estimating the precision of the score because it describes how likely it is for the obtained score to stray away from the true proficiency of the candidate.

The best way to decrease the SEM is by enlarging the sample that is, using more items on the examination. Thus, medical licensure and certification examinations tend to be long so as to minimize measurement error. This is especially true in the medical profession because the candidates are already highly selected by medical schools and other

examinations which decreases the variance in their scores, making it harder to tell them apart.
The precision of scores at the MPS is particularly important because it may result in wrong decisions by passing some who should have failed and failing some who should have passed. This misclassification error is directly related to the SEM, as well as other lesser factors. Thus we must have standards that control misclassification and keep it down to an acceptable level.[19]

The National Council on Measurement in Education (NCME) and the American Psychology Association (APA) require that testing agencies report the SEM at the cut score as a best practice.[19] This requirement is also required by agencies that accredit certification programs, such as the NCCA. This requirement gives the user the ability to evaluate the precision of the score on which decisions are made.

One simple example of the importance of precision in the scores compares two examinations: Examination A has a SEM of 2% and Examination B has a SEM of 10%. A candidate who has a true ability of 60%, will obtain a score that is likely to be between 58% and 62% on examination A, whereas that candidate will obtain a score of anywhere between 50% and 70% on Examination B. Thus, a score on Examination A is a much better (more precise) estimate of the candidate's true ability than a score on Examination B.

The SEM is related to the reliability of the scores, thus, examinations with higher reliability generally tend to produce scores with smaller average SEMs. However, reliability is not the only factor affecting SEM at the MPS. Targeting the difficulty of the items on the examination is important, because items yield maximum information when the ability of the candidate is near the target difficulty of the item.

To illustrate this point, consider the extreme case of a very difficult item to which only 1% of candidates respond correctly. This item does not give us differential information about the vast majority of candidates, since they all gave incorrect responses. The same is true of extremely easy items: they don't give us information about most of the candidates. The most informative items are ones that have their difficulty correspond to the ability of the group of interest. See Figure 1.

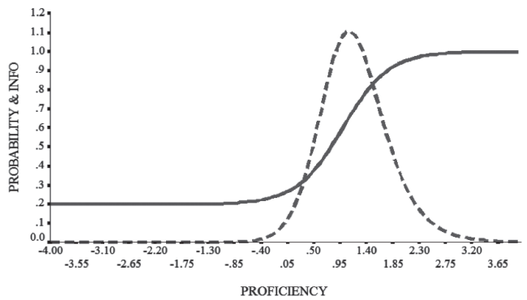

Figure 1. Characteristic and Information Curves of an Item

Figure 1 describes the probability of giving a correct response as a function of a candidate's ability. The difficulty of the item is defined at the point of inflexion of this curve (approximately at .95). This characteristic of the item is at the heart of Item Response Theory (IRT). The dotted curve represents the information function of that item, and it peaks at the point of inflection (difficulty level) of the item. Thus, if we want maximum information about candidates near the Pass/Fail decision point, we must use items that have difficulties in that region.[20]

Standard III: Methodology

The methods used in setting passing standards for a certification/licensing examination play an important role in defining the quality of the standards. It must be stressed at the outset that all standards are arbitrary decisions. What varies from method to method is: the nature of the standard, who is making the decision and how the decisions are made. While the decisions on the standard are arbitrary, they should be based on careful consideration of (a systematic study of) the variables involved, such as the nature of the decision, the purpose of the examination and its difficulty, the population being examined, social and economic impact.[13]
Generally, standard-setting methods fall into two major categories: norm-referenced methods and criterion-referenced methods. Norm-referenced methods result in relative standards that describe the minimum required level of competency in terms of the performance of a certain reference group. In contrast, criterion-referenced standards describe the minimum requirement as a level of proficiency or as an amount of content, regardless of the percentage of candidates failed or passed.
While the focus of norm-based standards is usually candidate fail rates that of criterion- referenced standards is minimum skill level or amount of knowledge, with little or no regard to fail rates. In applying a fully criterion-referenced standard all candidates may pass if they have the required proficiency or none may pass if they don't. Thus, these standard are considered to be harsh by some authorities.[7]
To address the potential harshness of criterion-referenced standards, another class of methods arose, that took note of proficiency level as well as fail rates. These methods are called compromise methods. It is probably a prudent policy to adopt a criterion-referenced study and then to temper it by a compromise method.[9,10]
Norm-referenced Standards

As mentioned before, these are based on the comparative performance of a reference group. Examples of such standards:

- 20% of applying candidates must fail.
- The MPS is set at 1.0 standard deviations below the mean score of the reference group.
- The positions are given to the 17 candidates with the highest scores.

It is clear that the standards above depend on how proficient the reference group is. If a candidate takes the examination with a strong group they are less likely to pass than with a weaker group.
Since groups tend to vary a little in proficiency from year to year, the standard will correspondingly change with each group's performance.
Criterion-referenced Standards

These standards require a certain amount or level of proficiency. Examples of such standards:

- Candidates must know at least 70% of the content to pass.
- To pass, candidates must demonstrate proficiency in skills at level 6.
- Candidates must be able to respond correctly to items of difficulty 2.0 to 2.5 log its 80% of the time. In medical certification and licensure, most examinations test knowledge of the content and so standards are usually expressed in amount of content mastered. It is assumed that the examination represents a sample of the general content domain, and that the score on the examination represents how much of the content was mastered by the candidate. In some examinations, however, the complexity or difficulty of the skills performed is the measure represented by the score. Whatever the case may be, the standard requires a certain proficiency or ability level regardless of who is taking the examination.

The first drawback of criterion-referenced standards is their possible harshness and lack of regard to fail rates. That can be mitigated by adding other considerations when deciding on the standard, for example, impact and policies. In other words, the results of the criterion-based study may be used in conjunction with other considerations, rather than be applied alone.[10]

The other drawback of these standards is that they may not be consistent from administration to administration. In the first example above, a 70% score is required for passing. But if the difficulty of the examination varies, as it often does, this score will represent different levels of proficiency. A 70% score represents a higher level of proficiency on a hard test than a 70% score on an easy one. Thus, scores on different administrations must be equated.
The importance of equating scores cannot be overstated if a consistent standard is to be applied on several administrations.[19] Equating removes (controls for) the effect of difference in difficulty between examinations and yields scores that are equivalent and directly comparable across administrations. It has other advantages such as enabling comparison and tracking of performance of individuals and/or programs across time.
The National Board of Medical Examiners© (NBME©), for example, uses a criterion-referenced standard-setting procedure.[4,15,16,18] However, the results of the procedure are not used alone in determining the passing standards, but combined with other information. Information sources used by NBME© to set standards are:

- Analyses of trends in performance and medical instruction,
- surveys of groups of interest in the medical community (medical schools, associations, etc.), and
- results of content-based standard-setting studies

In conclusion, standard setting is a complex political process with many facets and aspects. Resulting standards must, on one hand, protect the interests of the public and the user of the services, and on the

other hand, protect the rights of the candidate in fairness, equity and transparency. This balance is sometimes difficult to achieve, as it entails meticulous planning and attention to technical, psychometric, political and social factors.

BIBLIOGRAPHY

1. Schumacher CF. Reliability, validity, and standard setting. In: Hubbard JP, editor. Measuring Medical Education. 2nd ed. Philadelphia: Lea & Febiger; 1978. p.59-71.
2. National Board of Medical Examiners. Part I Standard setting procedures for 1981. The National Board Examiner, 1981; 28(1).
3. Swanson DB, Case SM, Kelley PR, Lawley JL, Nungester RJ, Powell RD & Volle, RL. Phase in of the NBME comprehensive Part I examination. Academic Medicine 1991; 66:443-444.
4. National Board of Medical Examiners. Development of the comprehensive Part I and Part II examinations. The National Board Examiner, 1990; 37(2).
5. Nungester RJ, Dillon GF, Swanson DB, Orr, NA & Powell RD. Standard setting plans for the comprehensive Part I and Part II examinations. Academic Medicine, 1991; 66:429-433.
6. National Board of Medical Examiners. Standard Setting and Score reporting for the comprehensive Part I and Part II Examinations. The National Board Examiner, 1991;38(3).
7. Angoff, W. Scales, norms and equivalent scores. In: Thorndike RL, editor. Educational Measurement, 2nd ed. Washington: American Council on Education, 1971. p.508-600.
8. Swanson DB, Dillon GF, & Ross, LP. Setting Content-Based Standards for National Board Exams: Initial Research for the comprehensive Part I Examination. Academic Medicine, 1990; 65:S17-S18.
9. Hofstee, WKB. The case for compromise in educational selection and grading. In Anderson SB and Helmick JS, editors. On Educational Testing. San Francisco: Jossey-Bass, 1983. p.109-127.
10. Beuk CH. A method for researching a compromise between absolute and relative standards in examinations. Journal of Educational Measurement, 1984; 21(2):147-152.
11. Dillon, GF. The expectations of standard-setting judges. The CLEAR Exam Review, 1996; 7:22-26.
12. Plake BS. Setting Performance Standards for Professional Licensure and Certification. Applied Measurement in Education, 1997; 11(1):65-80.
13. Orr NA & Nungester RJ. Assessment of Constituency Opinion about NBME Examination Standards, Academic Medicine,1991; 66:465-470.
14. Hallock JA, Melnick DE, Thompson JN. The Step 2 Clinical Skills Examination. JAMA. 2006; 295:1123-1124.
15. Melnick DE, Dillon GF, Swanson DB. Medical licensing examinations in the United States. Journal of Dental Education. 2002;66:595- 599.
16. Dillon GF, Case SM, Melnick DE, Nungester RJ, Swanson DE. Setting standards on the United States Medical Licensing Examination. In: Evolving Assessment: Protecting the Human Dimension. Proceedings of the Eighth International Ottawa Conference, 1998. Philadelphia, PA: National Board of Medical Examiners. 2000:466-474
17. Clauser BE, Mee J, Margolis MJ. The effect of data format on integration of performance data into Angoff judgments. International Journal of Testing. 2013;13:65-85.
18. Melnick DE. Licensing examinations in North America: is external audit valuable? Medical Teacher. 2009;31:212-214
19. AERA, APA, NCME. Standards for Educational and Psychological Testing. AERA 2014.

Artículo Original

Análisis Psicométrico de los Exámenes de Habilitación para el Ejercicio Profesional aplicados por el CEAACES

Lopez Meza Andrés H.

Consejo de Evaluación, Acreditación y Aseguramiento de la Calidad de la Educación Superior (CEAACES), Quito, Ecuador.

DIRECCIÓN PARA CORRESPONDENCIA: Andres Lopez Meza, Germán Alemán E11-32 y Javier Arauz, Quito, Ecuador. andres.lopez.m@outlook.com; andres.lopez@ceaaces.gob.ec

RESUMEN

Se realizó una revisión bibliográfica de la teoría del análisis psicométrico de las preguntas (ítems), utilizada en los Exámenes de Habilitación para el Ejercicio Profesional aplicados por el Consejo de Evaluación, Acreditación y Aseguramiento de la Calidad de la Educación Superior (CEAACES) en el Ecuador. Se definieron las dos corrientes teórico-metodológicas más importantes, la teoría clásica de los test (TCT) y la teoría de respuesta al ítem (TRI) - modelo logístico de un parámetro (modelo de Rasch). También se describieron los indicadores que se utilizan para evaluar las características psicométricas de los reactivos de un test: habilidad, dificultad, discriminación, correlación punto biserial, alfa de Cronbach, error estándar de medida. Por último, se revisó brevemente el análisis de distractores de los reactivos.

Palabras clave: Análisis psicométrico de reactivos, TCT, TRI, indicadores psicométricos, análisis de distractores.

ABSTRACT

A bibliographic review of the theory of psychometric analysis of questions (items) was conducted, which is used in the Qualification Exams for Professional Practice applied by the Consejo de Evaluación, Acreditación y Aseguramiento de la Calidad de la Educación Superior (CEAACES) in Ecuador. The two most important theoretical-methodological models were defined, the classical test theory (TCT) and the item response theory (IRT) - the logistic model of one parameter (Rasch model). In addition, the indicators used to evaluate the psychometric characteristics of items of these kinds of tests were described: ability, difficulty, discrimination, biserial point correlation, Cronbach's alpha, standard measurement error. Finally, the analysis of items' distractors was briefly explained.

Key words: Psychometric analysis of items, TCT, TRI, psychometric indicators, analysis of distractors.

INTRODUCCIÓN

El Consejo de Evaluación, Acreditación y Aseguramiento de la Calidad de la Educación Superior (CEAACES), por norma constitucional y legal, es el órgano público competente para aplicar los Exámenes de Habilitación para el Ejercicio Profesional (EHEP) en aquellas carreras que pudieran comprometer el interés público, poniendo en riesgo esencialmente la vida, la salud y la seguridad de la ciudadanía.

Uno de los requisitos para adquirir la certificación profesional en las carreras denominadas de interés público, es el aprobar el EHEP; este examen determina si los sustentantes han adquirido los conocimientos y habilidades necesarios para ejercer la profesión.

El objetivo de este estudio es dar a conocer una revisión bibliográfica de la teoría del análisis psicométrico de las preguntas, utilizada en los Exámenes de Habilitación para el Ejercicio Profesional aplicados por el CEAACES en el Ecuador. Se definirán los elementos fundamentales del análisis psicométrico, la conceptualización de las dos corrientes teórico-metodológicas más importantes y se revisará el análisis de preguntas de manera general. También se presenta una revisión de los indicadores que se emplean para la evaluación psicométrica las preguntas.