Five steps to effective test development

Cinco passos para o desenvolvimento de testes eficazes

Cinco pasos esenciales para el desarrollo de pruebas efectivas

Mary Foertsch*

Abstract

Effective test development requires a systematic, detail-oriented approach based on sound theoretical educational measurement principles. This paper discusses test development procedures or steps that typically must be accomplished in the development of most medical school assessments. Following these steps for effective test development tends to maximize validity evidence for the intended test score interpretation. These five-steps are presented as a framework, which test developers may find useful in organizing their approach to the many tasks commonly associated with test development – starting with detailed planning in Step 1, carrying through to discussions of content definition and delineation, and to creating test items. Test development consists of a series of interrelated activities, many of which depend on some prior step or steps. Careful planning and execution of a detailed plan leads to tests that more validly measure examinee ability or achievement in the well-defined content domain of interest. Adherence to this plan provides validity evidence from multiple sources. Each of the stages in the planning and development of a test is important to good measurement.

Keywords: Assessment; Medical education; Test development.

Resumo

O desenvolvimento de um teste eficaz requer uma abordagem sistemática, orientada para detalhes com base em princípios educacionais teóricos sólidos. Este artigo discute os procedimentos de desenvolvimento de testes ou etapas, que normalmente devem ser realizadas no desenvolvimento da maioria das avaliações para a escola de medicina. Seguindo esses passos, o desenvolvimento de um teste eficaz tende a maximizar a evidência de validade para a interpretação da pontuação no teste a que se destina. São apresentados cinco passos com uma estrutura que os desenvolvedores de testes podem utilizar na organização da abordagem para as muitas tarefas comumente associadas com o desenvolvimento de testes – começando com um planejamento detalhado na etapa 1, continua com a definição de conteúdo e delimitação, e a criação dos itens do teste. O desenvolvimento do teste consiste em uma série de atividades inter-relacionadas, muitas das quais dependem de algum passo ou etapa anterior. O planejamento cuidadoso e a execução de um plano detalhado produz testes que melhor avaliam a habilidade do examinado para o conteúdo na área de interesse definido. A adesão a este plano fornece evidências de validade a partir de múltiplas fontes. Cada uma das etapas do planejamento e desenvolvimento do teste é importante para uma medida adequada.

Descritores: Avaliação; Educação médica; Desenvolvimento de testes.

Resumen

El desarrollo de pruebas eficaces requiere un enfoque sistemático y preciso sobre la base de principios de medición educativos teóricos sólidos. Este articulo trata los procedimientos de desarrollo de pruebas o los pasos que normalmente se deben cumplir en el desarrollo de la mayoría de las evaluaciones de las escuelas de medicina. El seguimiento de éstos para el desarrollo de pruebas efectivas tiende a maximizar la evidencia de validez para la interpretación del puntaje de la prueba prevista. Estos cinco pasos son presentados como un marco, del que los desarrolladores de la prueba pueden servirse para organizar su enfoque, con las numerosas tareas comúnmente asociadas con el desarrollo de pruebas, comenzando con planificación detallada en el Paso 1, llevando a cabo dis-

Revista HUPE, Rio de Janeiro, 2014;13(4):79-83
Recebido: 26/05/2014 | Aprovado: 24/09/2014
doi: 10.12957/rhupe.2014.13961

*Corresponding address:
Virginia Tech Carilion School of Medicine.
2 Riverside Circle 24016.
Roanoke, VA, United States.
E-mail: mafoertsch@carilionclinic.org

cusiones sobre la definición y delimitación de contenidos, y creando ítems para prueba. El desarrollo de pruebas consiste en una serie de actividades relacionadas entre si, muchas de las cuales dependen de algún paso previo o pasos. Una planificación cuidadosa y la ejecución de un plan detallado, conduce a pruebas que miden más la capacidad del examinado que el logro en la definición de contenidos del dominio de interés. La adhesión a este plan proporciona la evidencia de validez a partir de múltiples fuentes. Cada uno de las etapas en la planeamiento, planificación y desarrollo de una prueba es importante para una medición buena.

Palabras clave: Evaluación; Educación médica; Desarrollo de prueba.

Introduction

A number of test development issues must be considered when developing a test for medical students. Issues in the planning of the test include developing a test blueprint, defining the test characteristics, developing items, obtaining item statistics, and finalizing specifications and procedures. These steps are presented in a linear model or as a sequential timeline, from a discrete beginning to a final end point; however, in practice, many of these activities may occur simultaneously or the order of some of these steps may be modified. The five steps listed here are the prerequisite to other activities; for example, content definition must occur before items are written and tests are assembled, so the sequence of steps, although somewhat arbitrary, is meaningful. The five essential steps in test development are the topic of this paper.

1. Determine the traits that the test should measure in order to develop the *test blueprint*. This process varies depending on the subject matter being assessed. For example, developing the test blueprint for an algebra test is fairly simple. The domain of knowledge to be assessed is well defined and easily identified. A much more complex situation is the development of a test blueprint for a medical licensure exam. These types of tests generally require a job or task analysis to determine the appropriate domain of content that should be included in the test blueprint. The result is a content blueprint, which facilitates item development and defines how many items should be included from each content area.
2. Define the *test characteristics* that are desired. At a minimum, this should include test length, measurement precision, and the maximum and acceptable rates at which items can be used or exposed to examinees. Technical properties such as rules and procedures for item selection, scoring methods, and the characteristics of reported scores should also be specified.
3. Develop or identify the set of items that might comprise a suitable item pool. An extensive item-writing and pretesting effort is necessary.
4. Obtain item statistics (either classical or IRT–based) to facilitate the construction of a test. Compare the results of the field test to the desired characteristics. Adjust the test blueprint, testing procedures, or outcome expectations as necessary.
5. Finalize the specifications and procedures.

Steps 1 through 5 are discussed in more detail next.

Development

Step 1: the test blueprint

Test blueprints or specifications are the details that define what a test is designed to measure and how the measurement will be accomplished. The test blueprint specifies the type, format, and content characteristics for each item on the test. The options and flexibility are almost endless. The test can be comprised of discrete items, stimulus-based items (i.e., items with graphics or introductory material) or a combination of the two.[1] Items can allow for any means of responding, provided scoring is accurate and reliable. Finally, the content characteristics of the items and the proportion of the test devoted to each content domain are specified. This in turn specifies the constraints under which items are selected for inclusion in a test form.

Step 2: the test characteristics

The decisions made at this point in time are

driven by the impending uses of the test and practical and economic considerations. The test developer frequently encounters these complicating matters that are either confounded or in conflict with one another.[2] For example, test length depends on how precise or reliable the test needs to be. Item quality enters into the equation at this point because tests consisting of more discriminating items can afford to be shorter than those based on less discriminating items. Item exposure control constricts access to the most discriminating items, thereby complicating the matter. This is a function of how secure the test needs to be and how frequently the test will be administered using the same items. Also, practical considerations put a limit on the amount of time an examinee can reasonably be asked to devote to the test. Finally, economic concerns are significant if test administration costs depend on the amount of time the examinee is seated in front of the computer. Sorting out all of these competing priorities can be a difficult and often contentious process. However, by the end of the process, decisions are required to balance all of these factors so that a test that meets the desired characteristics is produced.

During the development of blueprints, the following questions should be considered:

1. How reliable must the test be?
2. What is the format of the test? If a test is to have a specific number of questions, how long must it be to reach the required level of reliability? If a test is variable in length, what are the minimum and maximum lengths and the required precision?
3. How large should the pool of items be and what should its composition be?
4. What item selection rules and procedures will be used?
5. How frequently will items be permitted to be administered to examinees? What methods will be used to protect item security?
6. How will the test be scored? How will scores be reported to examinees?

Step 3: the item pool

An item pool is a collection of test items that can be used to assemble or construct a test. The item pool should include sufficient numbers of items of satisfactory quality, appropriately targeted to medical students' ability.[3] Requirements for the item pool are related to the test purpose, the test delivery method, the measurement of statistical assumptions of the model used to obtain the item characteristics within the pool (i.e., the model used to describe the interaction of an examinee with an item), the test length, the frequency of test administrations, and security requirements.[4] In other words, the item pool should be developed with the size and composition best suited to the test's content and technical requirements; however, compromise of some sort on this ideal may be necessary.

Items in an item pool usually have been administered previously so that they have some data or item characteristics associated with them.[5] These are called *operational items*. For example, it is typical for each operational item within a pool to have a difficulty index or value, as well as some measure if the item's potential to discriminate between examinees. In addition, items within a pool are associated with certain content classifications or categories.

Large medical school testing programs routinely add items to the available pool through regular item writing and pretesting efforts. The operational items in the pool can be contrasted with the *pretest* or *tryout* items. Pretest items are usually administered on a trial basis during an operational exam and are not scored. Typically, pretest items are administered to fewer examinees than operational or scored items. Therefore, the statistics computed for the pretest items often suffer from larger standard errors of estimate than their operational counterparts and may not reliably represent the item's performance on an operational test form. Nonetheless, if item analyses show that items are performing adequately, they can subsequently be used operationally.

Step 4: the item statistics

Most item pools consist of items with content identification and statistical or psychometric characteristics. It is rare for an item in an item pool to have no data at all. Usually it is assumed that the statistical data have been obtained from similar examinee populations. This is especially true when the statistical characteristics consist

of more traditional p-values (i.e., the item's difficulty index or proportion of correct responses) and point-biserial correlation coefficients (i.e., the item's discrimination index).[6]

It is also assumed that the statistics represent the performance of the item free from context effects. Context effects may include such things as page breaks, test speededness, and item-response dependency.[7] Most of these effects are due to conditions that can occur. For example, an item whose stem or stimulus might appear in its entirety on one test form but appear broken or on two pages in another form has different characteristics in each case simply due to the page-break effects.

Similarly, items that have appeared on a test that is speeded (i.e., one in which not all examinees reach those items near the end of the test with enough time to be able to provide thoughtful responses that reflect their true ability) may not have statistics that accurately describe the item's performance on a test that is not speeded.

Finally, items that appear together in a test form in such a way that some items interact with others to modify the responses to those items may exhibit certain statistical characteristics that may not reflect the item's performance when it appears without the other items. In these situations, the items are said to be *dependent* in the sense that the correct response to an item may depend on other items in addition to the ability of the examinee. However, the item may not be administered to an examinee in the same order or with the items that appeared with it on the first test form. Thus, item characteristics again may not accurately describe its performance on the second test form.

Step 5: document the procedures

For medical school examinations, test content defining methods must be systematic, comprehensive, and defensible. For instance, a medical school may wish to develop an end-of-curriculum comprehensive test covering the content of a two-year curriculum, with a passing score on this test required to continue to a third year of medical education. In this example, rigorous and defensible methods of content definition and delineation of the content are required; all decisions on content, item formats, and methods of content selection become essential aspects of validity evidence.[8]

It is important to finalize decisions used to establish test blueprints and specifications, their rationale, and the evidence from the test developers to support the argument that the test specifications fairly represent the content domain of interest. The use of unbiased, systematic, well-documented methods to create test blueprints and specifications is advised, so the resulting examination can have a reasonable chance of fairly representing the universe of content.

Conclusion

These five steps of effective test development provide a structured, systematic process for creating effective medical school examinations. Test development consists of a series of interrelated activities, many of which depend on some prior step or steps. Careful planning and execution of a detailed plan leads to tests that more validly measure examinee ability or achievement in the well-defined content domain of interest. Adherence to this plan provides validity evidence from multiple sources. Each of the stages in the planning and development of a test is important to good measurement. However, as previously noted, differing requirements for a testing program tend to serve as competing goals, and compromise across these competing goals or targets is often necessary.

High-quality test development demands great attention to detail. Test validity evidence is increased or decreased as the attention to detail increases or decreases. Quality control methods and procedures must be utilized to ensure that the intended inferences from the test scores are achieved and that validity is maximized. Systematically following these five steps for effective test development helps to ensure maximum test validity evidence for the tests that we develop.

References

1. 1.Haladyna TM, Downing SM. Construct-irrelevant variance in high-stakes testing. Educational Measurement: Issues and Practice. 2004,23(1):17-27. http://dx.doi.org/10.1111/j.1745-3992.2004.tb00149.x

2. Juul DD, Foertsch M. Developing and scoring multiple-choice examinations. In: Aminoff M, Faulkner L., editors. The American Board of Psychiatry and Neurology: Looking Back and Moving Ahead. Wash-

ington, DC: American Psychiatric Publishing; 2011. p. 175-186.

3. Ebel, R. Essentials of Educational Measurement. 5th printing edition. Englewood-Cliffs, NJ: Prentice Hall Publishers; p.356-376

4. Crowe A, Dirks C, Wenderoth M. Biology in Bloom: Implementing Bloom's taxonomy to enhance student learning in biology. American Society of Cell Biology. 2008;7:368-381.

5. Moore D, Green J, Gallis H. Achieving the desired results and improved outcomes: integrating planning and assessment throughout learning activities. Journal of Continuing Education Health Professionals. 2009;29:1-5.

6. Shaw D, Young S. Revised guidelines for conducting item analyses of classroom tests. The Researcher, Spring. 2004:15-22.

7. Thompson, Nathan A. A Journal of Applied Testing Technology.2008;9(5):1-17.

8. Liaison Committee on Medical Education. Accreditation Standards. Functions and Structure of Medical School. (2011). Available at http://www.lcme.org/publications/2014_2015_functions_and_structure_june_2013.

Mary Foertsch

Student Assessment. Virginia Tech Carilion School of Medicine. Roanoke, VA, United States.