Journal of Contemporary Medical Education
www.scopemed.org
DOI: 10.5455/jcme.20170705091608

Short Communication

Using the Spearman-Brown prophecy formula to improve medical school examination quality

Kenneth Royal

Department of Clinical Sciences, North Carolina State University, North Carolina, USA

Address for correspondence:
Kenneth Royal, Department of Clinical Sciences, North Carolina State University, North Carolina, USA.
E-mail: kdroyal2@ncsu.edu

Received: April 03, 2017
Accepted: June 20, 2017
Published: August 15, 2017

ABSTRACT

Most measurement experts suggest a minimum reliability estimate of 0.60-0.70 is desirable for routine medical education assessments with low-to-moderate stakes. The most common recommendation for improving examination reliability is to increase examination length. Unfortunately, adjustments in examination length may result in unpredictable reliability estimation and involves a great deal of trial and error for educators. The Spearman-Brown prophecy formula may help medical educators accurately predict the effects of adding or removing items on reliability estimation and help educators make immediate improvements to the psychometric quality and functioning of their examinations.

KEY WORDS: Health professions education, medical education, psychometrics, reliability, score reporting, testing

INTRODUCTION

It is a common practice for medical schools to utilize online software programs for administering assessments, such as midterm and final examinations. These software programs often are used by educators themselves or by assessment specialists who administer assessments on behalf of the educators. In virtually all instances, score reports are generated on conclusion of an assessment. These reports provide a variety of student performance statistics and a number of psychometric quality indicators. Detailed indicators typically include item difficulty and discrimination values and frequency tables denoting item distractor performance. Global indicators typically include descriptive summary statistics (e.g., mean, standard deviation, and standard error of mean measures), interitem correlations, normalized scores (e.g., percentile ranks and z-scores) and estimates of score reliability (e.g., Kuder-Richardson 20 coefficient). Reference guides often are included to help educators interpret each of the reported values.

While information currently provided on most score reports is sufficient for making evaluative judgments about the quality and functioning of an examination, this diagnostic information is not particularly helpful for medical educators who learn examination score reliability estimates are low and want to make immediate improvements. Thus, the purpose of this article is to call attention to the Spearman-Brown prophecy formula, a formula for predicting score reliability based on adjustments to test length, and illustrates how use of this formula can help medical educators make immediate improvements to their examinations without wasting valuable time with trial-and-error modification efforts.

Reliability

Sufficiently reliable scores are a hallmark of a quality examination and a necessary component of validity evidence [1,2]. Although the topic of reliability has been discussed extensively in the literature, briefly stated, reliability refers to the extent to which scores are reproducible on repeated trials [3]. All measurements contain some error, but it is the extent to which error can be minimized that determines how reproducible, or reliable, a set of scores will be. When little error is present, scores are highly reliable; when more error is present, scores are less reliable.

Most measurement experts suggest a minimum reliability estimate of 0.60-0.70 is desirable for routine classroom assessments, in which the stakes are low-to-moderate for students [4]. Although this is a useful guideline for medical educators to aspire, many remain unaware of what specifically can be done to improve reliability estimation. The most common recommendations for improving score reliability include increasing the length of the examination, improving item quality, and improving item targeting. Much has been written about how to construct quality examination items, and most medical educators are quite familiar with these concepts. Less familiar to educators, however, is the notion of how adjusting an examination's length can affect reliability estimation.

Spearman-Brown Prophecy Formula

The Spearman-Brown prophecy formula was originally published in 1910 by Charles Spearman and William Brown as independent articles in The British Journal of Psychology [5,6]. Both authors presented a formula for predicting reliability when test length was altered, so both were equally attributed to creating this mathematical formula. The concept underpinning the formula is rooted in classical test theory (CTT), which remains the primary psychometric approach for examination scoring of most routine college and university classroom examinations. In short, CTT is based on the linear relationship $X = T + E$, where the observed score (X) is equal to the true score (T) plus random error (E). When examinees are presented items, the true score and error components cannot be separated, but the variance attributed to both can be estimated. Therefore, it is possible to calculate test reliability, which is the ratio of true score variance to observed score variance.

The Spearman-Brown formula can be expressed in the following equations [7]. Equation 1 illustrates the Spearman-Brown formula when predicting the reliability after the test length has been altered:

$$\alpha^{new} = \frac{m\alpha^{old}}{1+(m-1)\alpha^{old}} \quad (1)$$

Where,

α^{new} = The new reliability estimate after altering examination length;

α^{old} = The reliability estimate of the current examination; and

m = The new examination length divided by the old examination length.

Equation 2 illustrates how the formula can be rearranged to determine the number of items necessary to achieve a desired reliability level when the original reliability is known.

$$m = \frac{\alpha^{new}(1-\alpha^{old})}{\alpha^{old}(1-\alpha^{new})} \quad (2)$$

Where,

α^{new} = The new reliability estimate after altering the examination length;

α^{old} = The reliability estimate of the current examination; and

m = The new examination length divided by the old examination length.

The critical assumption of the Spearman-Brown formula involves the use of new items with comparable psychometric properties. More specifically, new items should approximate the same level of difficulty and must measure the same construct to eliminate construct irrelevance variance. When items of comparable quality are added, the Spearman-Brown formula will predict an accurate reliability estimate given the change in test length. However, if lesser quality items are added, the formula will overestimate reliability; and if better quality items are added, the formula will underestimate reliability.

Recommendations

While it may be helpful to provide some medical educators with algebraic formulas for calculating various psychometric measures and indicators, most educators would likely prefer quick and easy calculations through an online calculator. Numerous worksheets and calculators are available online for calculating the Spearman-Brown prophecy formula, as well as working examples and demonstrations [8]. Medical educators should consult these resources to understand how to predict score reliability given examination modifications. Another recommendation is for examination software programs to include the Spearman-Brown prophecy formula as part of their software features. Programming this code into these software packages is not difficult and could easily be provided with sufficient requests/demands from customers.

Additional Considerations

A review of item statistics typically provides the diagnostic information necessary to identify problematic items. On identification, items may be revised (or removed) that contain bias, lack effective distractors, do not discriminate well, and so on. Each of these types of improvements can be made immediately and before the next administration of the examination. Making immediate improvements to reliability estimation, however, are generally more difficult and may require some guesswork and trial and error (often spanning years) without the assistance of the Spearman-Brown prophecy formula. The Spearman-Brown prophecy formula can help medical educators both immediately and accurately predict how reliability estimation will change if adjustments to examination length are made.

As noted previously, increasing examination length will increase reliability estimates. However, for the Spearman-Brown formula to accurately predict reliability, new items appearing on the examination must be of comparable quality to the existing items. Ideally, instructors could extend their examinations as necessary to achieve the minimum reliability values of 0.60 to 0.70 (depending on the stakes for examinees), but this is not always practical. Time constraints, availability of additional items, and concerns of examinee fatigue are all factors that can affect the degree to which instructors are able to adjust the length of their examinations. As a general rule, instructors are encouraged to provide as many quality items as possible given the designated amount of time per item. For example, if an instructor administers an examination for the duration of 1 h and allows students 1 min and 30 s per item, then medical educators should plan to make full use of the entire hour by providing as many items as possible, which in this case would be 40 items. Maximizing the number of items is helpful not only for increasing score reliability but also for increasing the inferential stability of students' scores as standard errors decrease in size when more items are administered.

CONCLUSION

Most measurement experts suggest a minimum reliability estimate of 0.60-0.70 is desirable for routine classroom

assessments. The most common recommendation for improving examination score reliability is to extend examination length. Unfortunately, adjustments in examination length may result in unpredictable score reliability and involve a great deal of trial and error for educators. The Spearman-Brown prophecy formula may help medical educators accurately predict the effects of adding or removing items on score reliability, and help educators make immediate improvements to the psychometric quality and functioning of their examinations. Educators should seek online calculators and spreadsheets to calculate these estimates for now. However, consumers of examination software programs (e.g., faculty, testing managers, and assessment specialists) should repeatedly request vendors make a reliability prediction feature available in their programming to improve assessment efforts.

REFERENCES

1. Messick S. Validity. In Linn RL, edtor. Educational Measurement. 3rd ed. New York, NY: Macmillan Publishing Co, Inc.; 1989. p. 13-103.
2. American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association; 2014.
3. Traub RE, Rowley GL. NCME instructional module: Understanding reliability. Educ Meas 1991;10:37-45.
4. George D, Mallery P. IBM SPSS Statistics 21 Step by Step: A Simple Guide and Reference. 13th ed. Boston, MA: Pearson; 2013.
5. Spearman C. Correlation calculated with faulty data. Br J Psychol 1910;3:271-95.
6. Brown W. Some experimental results in the correlation of mental abilities. Br J Psychol 1910;3:295-322.
7. Wells CS, Wollack JA. An Instructor's Guide to Understanding Test Reliability; 2003. Available from: http://www.testing.wisc.edu/Reliability.pdf. [Last accessed on 2017 Mar 21].
8. Royal KD, Hedgpeth MW. Balancing test length with sufficiently reliable scores. Educ Med J 2015;7. DOI: 10.5959/eimj.v7i1.321.

Source of Support: Nil, Conflict of Interest: None declared.