Education in Medicine Journal ISSN 2180-1932

INNOVATIVE IDEAS
Volume 7 Issue 1 2015
DOI: 10.5959/eimj.v7i1.321
www.eduimed.com

Balancing Test Length with Sufficiently Reliable Scores.

Kenneth D. Royal[1], **Mari-Wells Hedgpeth**[2]

[1]Department of Clinical Sciences, North Carolina State University, USA. [2]Office of Medical Education, University of North Carolina at Chapel Hill, USA.

ARTICLE INFO

Received	: 08/10/2014
Accepted	: 07/01/2015
Published	: 25/03/2015

KEYWORD

Psychometrics
Reliability
Testing
Medical Education
Assessment

ABSTRACT

One of the most important indicators of a quality examination is the reliability of the scores. In low to moderate stakes scenarios it is desirable for scores to achieve a minimum estimate of .70. Unfortunately, for many medical educators it is difficult to attain this minimum threshold for “acceptable” statistical reproducibility. A common approach is to include additional items to an exam, but this process can become cumbersome and misguided without clear direction. Fortunately, the Spearman-Brown Prophecy formula can help medical educators identify approximately how many additional items will be necessary to achieve a minimum reliability estimate of .70. This article describes a case in which we were presented with a less than desirable reliability estimate from a previous medical school examination, utilized the Spearman-Brown Prophecy formula, and was able to achieve the minimum estimate of .70 on the next iteration of the examination. We encourage others to make use of this technique rarely utilized outside of the psychometric arena as well.

CORRESPONDING AUTHOR: Kenneth D. Royal, Ph.D. Department of Clinical Sciences, NC State University, 1060 William Moore Dr. Raleigh, NC 27695. Email: kdroyal2@ncsu.edu

Introduction

One of the most important indicators of a quality examination is the reliability of the scores. Generally speaking, reliability refers to the extent to which the exam scores are statistically reproducible over repeated trials (1). In low to moderate stakes scenarios it is desirable for scores to achieve a minimum estimate of .70 (2). This is generally true regardless of the type of reliability estimate chosen (e.g., KR-20, Cronbach’s alpha, etc.). Psychometricians have long known that factors such as sample homogeneity, item difficulty, number of items, and the conditions under which exams are administered can impact reliability estimation (3). Despite one’s best efforts to standardize the administration of the exam and ensure only good psychometrically functioning items appear on an examination, reliability estimation can remain a bit unpredictable. In situations in which reliability estimates are lower than desired, test constructors may want to include additional items to potentially increase this estimate (It should be noted that any additional item must also be of sufficient psychometric quality). The problem, however, is one cannot pursue the minimum recommended estimate of .70 to such a degree that it results in an exam that is unduly long or arduous for test-takers. A healthy balance must be struck. Fortunately, there is a technique that can essentially predict how many additional items will be necessary to achieve a given reliability estimate.

Method

Setting

At the University of North Carolina at Chapel Hill, all medical school examinations are conducted online in a standardized format. All students, except for those with documented disabilities, are given the same amount of time to complete the examination. Proctors ensure the integrity of the examination scores by monitoring students throughout the examination process in large classrooms. Software programs that provide locked-down browsers and record students' monitors provide additional assurance that score results are trustworthy.

Previous Examination

A total of 180 students completed an examination based on basic science content during the 2012-2013 academic year. The examination consisted of 40 multiple-choice items, each with 5 response options. Psychometric results indicated each of the items discriminated sufficiently well, and there was no empirical evidence available to suggest any items should be removed.

The only discernible psychometric flaw was the examination possessed a Cronbach's alpha reliability estimate of .60. A value of .70 was necessary in order for the scores to be considered "acceptable" with regard to statistical reproducibility.

Procedures

A decision was made to include additional items on the examination, but the question of how many items to add in order to improve reliability estimation without burdening students with a very lengthy examination remained.

In order to get a reasonable estimate of how many additional (quality) items were necessary, we used the Spearman-Brown Prophecy formula (4). The formula can be expressed as:

$T = C * R_T * (1\text{-}R_C) / (1\text{-}R_T) * R_C$, where T = target number of items, R_T = target reliability, C = current number of items, and R_C = current reliability.

Values from the previous year's examination were substituted into the formula, which resulted in the following equation:

$T =40 * 0.7 * (1\text{-}0.6) / (1\text{-}0.7) * 0.6$.

Results indicated a total of about 62 items would be necessary to achieve a reliability estimate of .70.

Results

After including 23 additional items to the examination, a new examination consisting of 63 items was administered to a new cohort of students (n = 180). The Cronbach's alpha reliability estimate for the examination was .70. Each of the additional 23 items were evaluated for psychometric quality, and each presented evidence of adequate discrimination and difficulty.

Discussion

The Spearman-Brown Prophecy formula could predict within a reasonably precise manner the number of additional items necessary to reach our desired level of reliability. Additionally, the formula can be rearranged to predict reliability based on changes to the number of items. This formula would be expressed as:

$R_T = T * R_C / C * (1\text{-}R_C) + T * R_C$.

Although this formula was originally introduced more than a century ago, it is seldom used outside of the very narrow psychometrics arena. We have found this formula to be very useful in helping discern the impact the number of items will have on our reliability coefficients for our medical student examinations. Despite the age of this formula, it can still serve as an innovative tool for constructing robust medical school examinations. We encourage others in medical education to explore the use of this classic, but rarely utilized technique as well.

The only caveat to the use of the Spearman-Brown formula is the formula requires additional items to be of comparable quality to items existing on the current examination. If lesser quality items are added, the formula will overestimate reliability. If better quality items are added, the formula will underestimate reliability. As a general rule of thumb, it is best to ensure all examination items of sufficient psychometric quality before applying the Spearman-Brown formula, as inconsistencies could affect the precision of the formula.

Conclusion

Although the Spearman-Brown Prophecy formula was derived more than a century ago, its usage outside the psychometric arena has been incredibly limited. We explored the predictive ability of this formula with a medical school exam and found the formula correctly predicted the number of additional items we would need to include to reach a reliability estimate of .70. Given most medical school examinations are at least moderate-stakes for examinees, it is critical that medical educators produce examinations with desirable psychometric properties, such as sufficient reliability (statistical reproducibility). Reliability is particularly important as it is an indicator of score validity (5), and a necessary component of a legally defensible examination. We encourage others to use the Spearman-Brown Prophecy formula as needed to develop the most robust examinations possible.

Reference

1. Albanese Royal KD. Understanding reliability in higher education student learning outcomes assessment. Qual Approaches High Educ. 2011;2(2):8-15.
2. George D, Mallery P. *IBM SPSS Statistics 21 Step by Step: A Simple Guide and Reference* (13th ed). Pearson; 2013.
3. Royal KD, Puffer JC. The reliability of American Board of Family Medicine examinations: Implications for test-takers. J Am Board Fam Med. 2012;25(1):131-133.
4. Spearman C. Correlation calculated from faulty data. Br J Psychol. 1910;3:271–295.
5. Messick S. Validity. In Linn RL ed. *Educational measurement* (3rd ed). New York, NY: Macmillan Publishing Co, Inc; 1989;13-103.