Journal of Modern Applied Statistical Methods

Volume 7 | Issue 1 Article 12

5-1-2008

Estimating How Many Observations are Needed to Obtain a Required Level of Reliability

David A. Walker
Northern Illinois University, dawalker@niu.edu

Journal of Modern Applied Statistical Methods
May, 2008, Vol. 7, No. 1, 152-157

1538 – 9472/08/$95.00

Estimating How Many Observations are Needed to Obtain a Required Level of Reliability

David A. Walker
Northern Illinois University

This article provides a detailed table containing estimations of how many observations are needed to obtain an increased reliability coefficient for situations such as observational data collection in the classroom. A SPSS program is provided for users to analyze situations where an initial reliability value is obtained and the user wants to determine how many more observations are needed to reach a required level of reliability.

Key words: Spearman-Brown, reliability, observations.

Introduction

The Spearman-Brown Prophecy Formula (SBPF) is often employed to estimate split-half reliability, as a function of internal consistency, of the variability of scores split on a composite test, and based on the assumptions that the two halves of a test have equal variance parameters and are consistent in content (Kristof, 1974; Zimmerman, 1970). An application of the Spearman-Brown formula is to estimate how many items need to be added to a test to obtain a specified level of reliability (Burnett, 1974). Li and Wainer (1997) noted that the Spearman-Brown formula's principal use has been to obtain, "… the reliability coefficient for a composite measurement as the sum of *n* individual measurements…" (p. 479). Its calculation is used as a function of estimating the score reliability of lengthened or shortened tests. The general formula for the SBPF, given by Krathwohl (1993), is expressed as

$$r = (kr) / [1 + (k - 1)r] \qquad (1)$$

David Walker is an Associate Professor of Educational Research. His Interests include structural equation modeling, effect sizes, factor analysis, predictive discriminant analysis, and bootstrapping. Email him at dawalker@niu.edu.

where, k = the ratio of items in the new test to those in the original form; r = the average of the sample correlations among individual measures. A simplified version of the Spearman-Brown Prophecy Formula, noted by Charter (2001) can be expressed as:

$$r_{kk} = 2r_{12} / (1 + r_{12}) \qquad (2)$$

where, r_{xy} = the correlation between the two halves of a test;

As an extension of the use of the Spearman-Brown formula, classroom observations and raters' judgments have been added to this application of it by expanding its use to situations for estimating the reliability of pooled judgments or observations (cf. Blok, 1985; Jenkins, Bausell, & Magoon, 1972). Hartmann (1976) describes these instances as N = 1 designs, which are "specifically relevant to reliability assessment [and] involve sessions, observers, and trails (multiple brief observation periods) within sessions" (p. 844).

There are various sources of error affiliated with classroom observational data or pooled judgments. For example, but not all inclusive, error can be derived from the length of an observation, with shorter observations a prevailing source of error; from a lack of equivalence between raters, which is often difficult to obtain with consecutive observational tasks; from observational processes that may cause variability among raters; from inter-rater disagreement; or from large deviations in performers' performances across observational points (Blok, 1985; Hartmann, 1976; McGaw,

Wardrop, & Bunda, 1972; Rogosa & Ghandour, 1991; Rowley, 1978).

To minimize sources of error, Rowley (1978) found that the pooling together of observational periods so that they occur more frequently, instead of prolonged observations, is more beneficial to reliability, "Reliability will be enhanced by a more representative sampling of occasions, and this is best achieved by using a larger number of shorter observation periods" (p. 172). Medley and Mitzel (1963) determined that an increase in congruent observational periods, but not an increase in observers, could lessen measurement error. Finally, Meehl (1999) found that judges' ratings pertaining to a common objective, or pooling their judgments, can increase reliability and is a beneficial technique: "If we have the judgments of only a few scientists (rating a batch of theories of single experiment), we can estimate the reliability of a larger pooled judgment via the Spearman-Brown Prophecy Formula…. to predict the boosted reliability of a lengthened mental test, [it] has turned out to be quite accurate when the elements are not test items but human judgments." (p. 292)

Purpose

The purpose of this study is to provide researchers with a detailed table containing estimations of how many observations are needed to obtain an increased reliability coefficient for situations such as observational data collection in the classroom. As well, SPSS (Statistical Package for the Social Sciences) syntax is provided in Appendix A for users to create Table 1 or analyze other situations where an initial reliability value is obtained and the user wants to determine how many more observations are needed to reach a required level of reliability.

Results

As can be seen from Table 1, when the initial reliability from 1 observation is very low, ranging from .100 to .200, it would require between 81 (i.e., r = .100) to 36 (i.e., r = .200) observations to increase reliability to a level of .900, respectively. Further review of Table 1 indicates that as the initial reliability measure increases into the moderate range (e.g., ≥ .600), the number of observations needed to enhance reliability would decrease, which is to be expected. Further, the data in Table 1 can be used as a scale by researchers involved in observational types of studies to determine, based on a preliminary measure of reliability, how many more observations would be required to reach a required level of reliability.

Usage Example

Assuming that many of the potential sources of error noted previously with use of this form of the Spearman-Brown formula were addressed and the user understood the tenets of reliability in terms of employment of observational protocols, calculation, and the interpretation of results; and the contextual uses of the SBPF versus coefficient alpha, for instance, for certain applications (cf. Charter, 2001; Martin, 1977), which admittedly may not be the case in every situation, the use of the Spearman-Brown formula to estimate the reliability of pooled judgments or observations may be warranted. For example, if a college or university-level researcher were conducting classroom-based research and performed an initial observation in a class that lasted for 20 minutes and obtained a score reliability estimate of .600 derived from the protocol used in the observation (i.e., the left column of Table 1), to increase the reliability to a desired level of .800 (i.e., the center column of Table 1), the researcher would need 3 more congruent observational periods (i.e., the right column of Table 1).

Rowley (1978) demonstrated this concept in a much more truncated example than Table 1, where it was determined that "… we may observe that a reliability of .176 obtained from one 10-minute visit could be increased to … .516 by making five times as many visits" (p. 170). Rowley's example can be replicated in the syntax in Appendix A by entering in the initial reliability level of .176 in the left column between the BEGIN DATA and END DATA field, putting in the desired reliability level of .516 in the right column of the same field, and then running the program, which will produce the number of observational periods needed of 5.

Table 1. The Number of Observations Needed to Obtain an Increased Reliability Coefficient

Initial Reliability Estimate	Increased Reliability Estimate	Observations Needed		Initial Reliability Estimate	Increased Reliability Estimate	Observations Needed
0.1	0.2	2		0.2	0.3	2
0.1	0.25	3		0.2	0.35	2
0.1	0.3	4		0.2	0.4	3
0.1	0.35	5		0.2	0.45	3
0.1	0.4	6		0.2	0.5	4
0.1	0.45	7		0.2	0.55	5
0.1	0.5	9		0.2	0.6	6
0.1	0.55	11		0.2	0.65	7
0.1	0.6	13		0.2	0.7	9
0.1	0.65	17		0.2	0.75	12
0.1	0.7	21		0.2	0.8	16
0.1	0.75	27		0.2	0.85	23
0.1	0.8	36		0.2	0.9	36
0.1	0.85	51		0.2	0.95	76
0.1	0.9	81		0.25	0.35	2
0.1	0.95	171		0.25	0.4	2
0.15	0.25	2		0.25	0.45	2
0.15	0.3	2		0.25	0.5	3
0.15	0.35	3		0.25	0.55	4
0.15	0.4	4		0.25	0.6	4
0.15	0.45	5		0.25	0.65	6
0.15	0.5	6		0.25	0.7	7
0.15	0.55	7		0.25	0.75	9
0.15	0.6	9		0.25	0.8	12
0.15	0.65	11		0.25	0.85	17
0.15	0.7	13		0.25	0.9	27
0.15	0.75	17		0.25	0.95	57
0.15	0.8	23		0.3	0.4	2
0.15	0.85	32		0.3	0.45	2
0.15	0.9	51		0.3	0.5	2
0.15	0.95	108		0.3	0.55	3

Table 1 (cont'). The Number of Observations Needed to Obtain an Increased Reliability Coefficient

Initial Reliability Estimate	Increased Reliability Estimate	Observations Needed		Initial Reliability Estimate	Increased Reliability Estimate	Observations Needed
0.3	0.6	4		0.45	0.8	5
0.3	0.65	4		0.45	0.85	7
0.3	0.7	5		0.45	0.9	11
0.3	0.75	7		0.45	0.95	23
0.3	0.8	9		0.5	0.6	1
0.3	0.85	13		0.5	0.65	2
0.3	0.9	21		0.5	0.7	2
0.3	0.95	44		0.5	0.75	3
0.35	0.45	2		0.5	0.8	4
0.35	0.5	2		0.5	0.85	6
0.35	0.55	2		0.5	0.9	9
0.35	0.6	3		0.5	0.95	19
0.35	0.65	3		0.55	0.65	2
0.35	0.7	4		0.55	0.7	2
0.35	0.75	6		0.55	0.75	2
0.35	0.8	7		0.55	0.8	3
0.35	0.85	11		0.55	0.85	5
0.35	0.9	17		0.55	0.9	7
0.35	0.95	35		0.55	0.95	16
0.4	0.5	1		0.6	0.7	2
0.4	0.55	2		0.6	0.75	2
0.4	0.6	2		0.6	0.8	3
0.4	0.65	3		0.6	0.85	4
0.4	0.7	3		0.6	0.9	6
0.4	0.75	4		0.6	0.95	13
0.4	0.8	6		0.65	0.75	2
0.4	0.85	8		0.65	0.8	2
0.4	0.9	14		0.65	0.85	3
0.4	0.95	28		0.65	0.9	5
0.45	0.55	1		0.65	0.95	10
0.45	0.6	2		0.7	0.8	2
0.45	0.65	2		0.7	0.85	2
0.45	0.7	3		0.7	0.9	4
0.45	0.75	4		0.7	0.95	8

Table 1 (cont'). The Number of Observations Needed to Obtain an Increased Reliability Coefficient

Initial Reliability Estimate	Increased Reliability Estimate	Observations Needed		Initial Reliability Estimate	Increased Reliability Estimate	Observations Needed
0.75	0.85	2		0.8	0.95	5
0.75	0.9	3		0.85	0.9	2
0.75	0.95	6		0.85	0.95	3
0.8	0.9	2		0.9	0.95	2

This article provides a detailed table containing estimations of how many observations are needed to obtain an increased reliability coefficient for situations such as observational data collection in the classroom. As well, SPSS syntax is provided for users to analyze situations where an initial reliability value is obtained and the user wants to determine how many more observations are needed to reach a required level of reliability.

This article could be of use to researchers who carry-out school-based research studies, those who conduct classroom-based observations, for example, of student teachers, student engagement, leadership capacity, or those engaged in decision-making studies related to a specified criterion. Thus, the merit in the use of the program in Appendix A or Table 1 is to assist researchers with an easily understood method to determine if the initial score reliability from an observational protocol used in a classroom to measure a particular trait or performance is on target or are further, congruent observational periods needed to reach a desired level of score reliability.

References

Blok, H. (1985). Estimating the reliability, validity, and invalidity of essay ratings. *Journal of Educational Measurement, 22*, 41-52.

Burnett, J. D. (1974). Parallel measurements and the Spearman-Brown Formula. *Educational and Psychological Measurement, 34*, 785-788.

Charter, R. A. (2001). It is time to bury the Spearman-Brown "prophecy" formula for some common applications. Educational and Psychological Measurement, 61, 690-696.

Hartmann, D. P. (1976). Some restrictions in the application of the Spearman-Brown Prophesy Formula to obervational data. *Educational and Psychological Measurement, 36*, 843-845.

Jenkins, J. R., Bausell, R. B., & Magoon, A. J. (1972). Selection of prose material for testing. *Journal of Educational Measurement, 9*, 97-103.

Krathwohl, D. R. (1993). Methods of educational and social science research: An integrated approach. White Plains, NY: Longman.

Kristof, W. (1974). Estimation of reliability and true score variance from a split of a test into three arbitrary parts. *Psychometrika, 39*, 491-499.

Li, H., & Wainer, H. (1997). Toward a coherent view of reliability in test theory. *Journal of Educational and Behavioral Statistics, 22*, 478-484.

Martin, J. (1977). The development and use of classroom observation instruments. *Canadian Journal of Education, 2*, 43-54.

McGaw, B., Wardrop, J. L., & Bunda, M. A. (1972). Classroom observation schemes: Where are the errors? *American Educational Research Journal, 9*, 13-27.

Medley, D. M., & Mitzel, H. E. (1963). Measuring classroom behavior by systematic observation. In N.L. Gage (Ed.), *Handbook of research on teaching* (pp. 267-269). Chicago: Rand McNally.

Meehl, P. E. (1999). How to weight scientists' probabilities is not a big problem: Comment on Barnes. *British Journal for the Philosophy of Science, 50*, 283-295.

Rogosa, D., & Ghandour, G. (1991). Statistical models for behavioral observations. *Journal of Educational Statistics, 16*, 157-252.

Rowley, G. L. (1978). The relationship of reliability in classroom research to the amount of observation: An extension of the Spearman-Brown formula. *Journal of Educational Measurement, 15*, 165-180.

Zimmerman, D. W. (1970). Variability of test scores and the split-half reliability coefficient. *Educational and Psychological Measurement, 30*, 259-266.

Appendix A. Syntax for Estimation of How Many Observations are Needed to Obtain an Increased Reliability Coefficient.

```
**************************************************************************
Author: David A. Walker (2008), dawalker@niu.edu,  Northern Illinois University
**************************************************************************

DATA LIST LIST/ r REST (2F9.3).

***NOTE: As the first number between BEGIN DATA and END DATA, put your initial score
reliability and then as the second number, put the estimated, increased score reliability that you
would like to achieve***.

BEGIN DATA
.176 .516
END DATA.
COMPUTE OBS = (REST*(1-r)/(r*(1-REST))).
EXECUTE.
FORMAT OBS (F8.0).
VARIABLE LABELS r 'Single Observation Reliability'/REST 'Estimated, Boosted
Reliability'/OBS 'The Number of Observations Needed to Equal an Estimated, Boosted
Reliability'/.
REPORT FORMAT=LIST AUTOMATIC ALIGN (LEFT)
MARGINS (*,110)
  /VARIABLES= r REST OBS
  /TITLE "Estimation of How Many Observations are Needed to Obtain an Increased Reliability
Coefficient".
```