Journal of Clinical Epidemiology 85 (2017) 45–49

Journal of Clinical Epidemiology

Spearman–Brown prophecy formula and Cronbach's alpha: different faces of reliability and opportunities for new applications

Henrica C.W. de Vet[a,*], Lidwine B. Mokkink[a], David G. Mosmuller[b], Caroline B. Terwee[a]

[a]*Department of Epidemiology and Biostatistics, Amsterdam Public Health Research Institute, VU University Medical Center, De Boelelaan 1089A, Amsterdam 1081HV, The Netherlands*

[b]*Department of Plastic Surgery, VU University Medical Center, De Boelelaan 1117, Amsterdam, 1081 HV, The Netherlands*

Accepted 13 January 2017; Published online 22 March 2017

Abstract

Objectives: There are similarities between the different forms of reliability, such as internal consistency (internal reliability) and interrater and intrarater reliability. Reliability coefficients that are based on classical test theory can be expressed as intraclass correlation coefficients (ICCs), such as Cronbach's alpha. The Spearman–Brown prophecy formula (SB formula) is used to calculate the reliability when the number of items in a questionnaire is changed. This paper aims to increase insight into reliability studies by pointing to the assumptions of reliability coefficients, similarities between various coefficients, and the subsequent new applications of reliability coefficients.

Design, Settings and Results: The origin and assumptions of Cronbach's alpha and the SB formula are discussed. Cronbach's alpha is written as an ICC formula, using the well-known property that taking the average value of a number of ratings increases the reliability of a measurement. We illustrate with an example that the ICC formulas for average measurements of multiple raters and the SB formula give similar results. This implies that the SB formula can be used to decide on the number of measurements to be averaged and thus on the number of raters required, for obtaining measurements with acceptable reliability, even if the variance components of the ICC formula are not known. Using the same example, we illustrate the principle of "Cronbach's alpha if item deleted" to decide on the poorest performing raters in a set of raters.

Conclusion: These applications have different assumptions: the principle of "Cronbach's alpha if item deleted" is based on the assumption of a fixed set of items/raters and the SB formula is based on the assumption of random raters. The example also emphasizes the need for more raters in the design of the reliability study to obtain a robust estimation of reliability.

Keywords: Reliability; Intraclass correlation coefficient; Spearman–Brown prophecy formula; Cronbach's alpha; Classical test theory

1. Introduction

In the COSMIN taxonomy, the domain of reliability consists of three measurement properties: internal consistency, reliability, and measurement error [1]. In a previous paper, we focused on the similarities and differences between reliability and measurement error [2]. In this paper, we will explain the conceptual similarities between the measurement properties internal consistency and reliability. Deepening our understanding of the theory of reliability will provide opportunities for new applications of well-known reliability coefficients.

The COSMIN initiative defined the domain of reliability as the extent to which scores for persons (or patients) who have not changed are the same for repeated measurement under varying conditions [1]. Repeated measurements are, for example, measurements using different sets of items from the same questionnaire (internal consistency); measurements at different time points (test–retest); measurements by different persons on the same occasion (interrater) or by the same persons on different occasions (intrarater). "Raters" is used as an umbrella term for respondents, observers, assessors, etc., and "persons" as an umbrella term for patients, subjects, or objects that are targets of reliability assessments, such as images, biopsies, or photographs. The domain of reliability includes internal consistency, measurement error, and reliability, the latter being defined by COSMIN [1] as the proportion of the total variance in the

Conflict of interest: None of the authors has a conflict of interest.

* Corresponding author. Tel.: +31-20-4446014; fax: +31-20-4446775.
E-mail address: hcw.devet@vumc.nl (H.C.W. de Vet).

http://dx.doi.org/10.1016/j.jclinepi.2017.01.013
0895-4356/

What is new?

Key findings

- Cronbach's alpha and the Spearman–Brown prophecy formula have broader applications than for which they were originally developed.

What this adds to what was known?

- The principle of "Cronbach's alpha if items deleted" can be applied in a pilot study to select raters from a fixed set of raters to perform measurements in a large-scale study.
- The Spearman–Brown formula can be applied to decide on the number of measurements to be averaged to obtain acceptable reliability.
- Assumptions about random or fixed raters are important.

What is the implication and what should change now?

- The application of Cronbach's alpha and the Spearman–Brown formula in research and clinical practice can be extended when keeping assumptions about random or fixed raters in mind.

measurements which is due to "true" differences between patients, in accordance with the original definition by Lord and Novick [3] who defined reliability as the ratio of the true score variance and the total variance.

There are a number of similarities between the different forms of reliability. The basic assumption of all forms of reliability is that the person to be measured remains stable on the construct of interest and that differences between the repeated measurements are considered error variance. The objective is to capture a person's "true" score, and all measurements are accompanied by some error variance [3]. All reliability coefficients based on classical test theory can be expressed as intraclass correlation coefficients (ICCs), including Cronbach's alpha [4]. Cronbach's alpha is known as a measure of internal consistency used in the context of multi-item measurement instruments [5–7]. The Spearman–Brown prophecy formula (SB formula) is perhaps less well known to epidemiologists: it is used to predict reliability when the number of repeated measurements, for example, the numbers of items in a multi-item questionnaire, is changed [8,9]. The recognition of the similarities in reliability coefficients used to assess internal consistency and to assess reliability, concerning items, and raters, respectively, opens up the opportunity to share each other's applications.

We will start by explaining how ICCs are calculated and demonstrate that Cronbach's alpha can be written as an ICC [4]. Subsequently, we will introduce an example and show two applications which are based on the similarity between internal consistency and interrater reliability:

1. Application of the SB formula that was originally developed for predicting the reliability of scales depending on the number of items, for determining the number of raters required to obtain reliable measurements.
2. Application of the principle of "Cronbach's alpha if item deleted" to delete poor performing raters in a set of raters similar to deleting the poorest performing items in a questionnaire.

We will end with some remarks and assumptions regarding these applications and new lessons with regard to robustness of estimations of reliability.

2. Intraclass correlation coefficients

Knowledge of the formula of ICC will enhance the understanding of the similarities between different reliability coefficients, including Cronbach's alpha. All ICCs are based on a ratio of variances which can be derived from analysis of variance (ANOVA). In an ANOVA analysis, the total variance of the measurements can be split into true variance (i.e., the "true" differences between persons) and error variance. The ICC relates the true variance, that is, the variance due to the differences between persons (σ_p^2) to the total variance in the measurements, which comprises the true variance (σ_p^2) plus the error variance $(\sigma_{\text{error}}^2)$. The general formula for the ICC is

$$\text{ICC}_{\text{general}} = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_{\text{error}}^2} \quad (1)$$

There are various ICC formulas, as the composition of the error variance may differ. The appropriate formula depends on the type of errors one is interested in and the design of the study [4]. In this paper, we use a so-called complete design, the simplest design, where all persons are assessed by all raters or items. We will introduce only the ICC formulas that are necessary to understand the issues explained in this article. To this end, we will explain the distinction between random and fixed raters in a reliability study and present the formulas for calculating the reliability of single and average measurements.

Most studies use random samples from a population of interest. Raters can be considered either random or fixed. Random raters are viewed as a random sample of all possible raters. In statistical terms, a model with random persons and random raters is considered a random model (one of the options when calculating ICC in SPSS reliability analysis). In some situations, we are interested in a specific set of raters, for example, to assess the performance of all pathologists who work in a specific department. In that case, we consider the raters as fixed and the model is

called mixed because it contains a mixture of random factors (persons) and fixed factors (raters). In a two-way ANOVA, the error variance (σ^2_{error}) is split into a random error term ($\sigma^2_{\text{residual}}$) and a systematic error term (σ^2_r) that indicates the systematic differences among the raters. In the random model, both are included in the denominator of the ICC, and we use the term $\text{ICC}_{\text{agreement}}$. In the mixed model, only the random error term is included in the denominator and we use the term $\text{ICC}_{\text{consistency}}$ [2,4].

$$\text{ICC}_{\text{agreement}} = \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_r + \sigma^2_{\text{residual}}} \quad (2A)$$

$$\text{ICC}_{\text{consistency}} = \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_{\text{residual}}} \quad (2B)$$

The reliability can be increased by taking the average score of multiple measurements [4]. If, for example, the scores of three raters are averaged, the reliability of the average score will be higher than the reliability of each single score. An application of this idea in clinical practice is the way in which general practitioners measure blood pressure. Blood pressure measurements are known to be unreliable. A well-known strategy is to measure the blood pressure of a patient three times and take the average value. In the general ICC formula, the error variance (σ^2_{error}) is divided by a factor 3. In this way, the reliability of the blood pressure measurements is increased. The formulas are presented in Appendix A.1 at www.jclinepi.com.

The ICC coefficient representing Cronbach's alpha can be written as

$$\text{Cronbach's alpha} = \text{ICC}_{\text{consistency}} = \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_{\text{residual}}/k} \quad (3)$$

where k is the number of items [4]. Cronbach's alpha is represented as an $\text{ICC}_{\text{consistency}}$ because the items, the specific items in the questionnaire, are fixed. The reason that the error variance (i.e., $\sigma^2_{residual}$) is divided by the number of items is because we always use the total score of the questionnaire, being either a sum or an average of the k item scores. So instead of averaging the scores of three raters, we average the scores of a number of items. Note that formula 3 also illustrates why Cronbach's alpha becomes higher when the number of items increases: the error variance is divided by the number of items.

3. An example

In a study with 50 6-year-old children with complete and incomplete unilateral cleft lip and palate, the cosmetic result after surgical lip repair was assessed by judging symmetry of the nose and lip based on photographs using a program named SymNose [10]. Four raters scored the symmetry of, among other things, the front view of the nose on all photographs, and interrater reliability was assessed. Details of the design of the study and its results will be published elsewhere [11].

We performed a reliability analysis in SPSS calculating the $\text{ICC}_{\text{agreement}}$ for interrater reliability based on the result of these four raters. The output (presented in Appendix A.2 at www.jclinepi.com) shows several reliability measures: a value for Cronbach's alpha, an ICC value for single measures, and an ICC value for average measures. The ICC value for single measures ($\text{ICC}_{\text{agreement}} = 0.654$) represents the reliability of one score obtained by one rater. This is usually how measurements are done in clinical practice; for example, one physician scores the symmetry of the nose and lip on a photograph of one patient. Although we use four raters in the study and 50 patients (photographs), the ICC refers to the reliability of one score of one patient obtained from one rater. We needed the other raters and patients just to enable calculation of the reliability coefficient. The ICC value for "single measures" (0.654) is not very high which means that if another physician had scored the same patient, the score might be different. The ICC value for average measures ($\text{ICC}_{\text{agreement}} = 0.883$) is higher. This represents the reliability of the average score of the four raters for one patient. This ICC value for average measures would only apply if in clinical practice each photograph was rated by four raters (and their scores averaged), that is, comparable to the strategy used for blood pressure measurements. The value of Cronbach's alpha is almost identical to the ICC value for average measures (0.886 vs. 0.883, respectively). Cronbach's alpha also represents the reliability of the average score of the four raters. The difference is that reliability is now calculated as $\text{ICC}_{\text{consistency}}$ (formula 3) not taking into account the systematic error term (σ^2_r), which is very small in this example (see Appendix A.3 at www.jclinepi.com for formulas and calculations).

4. Two new applications

4.1. Using the Spearman–Brown prophecy formula for estimating the number of raters

The SB formula was originally developed independently by Spearman [8] and Brown [9] published in the same journal in 1910 and nowadays still a topic of interest [7]. The formula predicts the reliability of a questionnaire when using the subgroups of items (split in half, thirds, fourths, etc.) to examine internal consistency. They knew that an instrument with, for example, half the number of items underestimates the reliability of the full length test. The SB formula was used to predict reliability for another number of items assuming the average correlation (mean r) remains the same. It is important to note that the items should be randomly split into subgroups, which means that both the expected true score and the error variance are the same for each subgroup. In psychometric terms, the SB formula requires the condition of parallel tests [2,7]. This

condition is often assumed to be satisfied in interrater studies as the raters are estimating the same true scores, but only after the data are collected it can be determined if the condition of parallel tests holds.

In the formula

$$r_{\text{Spearman-Brown}} = \frac{nr}{1+(n-1)r} \quad (4)$$

n is the factor by which the number of items will be multiplied, and r is the reliability (internal consistency) of the questionnaire. For example, if the questionnaire is shortened by a factor 2, n will be 0.5, and if the questionnaire contains twice as many items, n will be 2. This is useful in the developmental phase of a questionnaire when internal consistency appears to be rather low because only a few items are included. In that case, the researchers could add items to improve the reliability of the questionnaire.

The SB formula can also be used for raters instead of items, that is, to predict interrater reliability when the number of raters changes. If the reliability found for a single measurement is low (as in our previous example with an $ICC_{agreement}$ of 0.654), one can calculate how many scores need to be averaged, so how many raters would be required, to achieve acceptable reliability (e.g., ICC $>$ 0.70). Fig. 1 shows how the value of a reliability coefficient for a single rater corresponds to reliability coefficients for the average value of the score of two, three, or four raters. Note that the SB formula can be applied to both $ICC_{agreement}$ and $ICC_{consistency}$. Appendix A.4 at www.jclinepi.com contains the calculations.

With the SB formula, we can also calculate the reliability for the average score of two or three raters, which for the $ICC_{agreement}$ results in 0.791 and 0.850, respectively. A reliability of more than 0.80 for the symmetry assessment requires the average score of at least three raters. The output of the SPSS reliability analysis does not present the average scores of two or three observers, but we can calculate these values using the formulas in Appendix A.2 at www.jclinepi.com. The main advantage of using the SB formula is that we can even calculate the $r_{Spearman-Brown}$ when we do not have direct access to the raw data and we therefore do not know the values of all variance components. So, if an ICC of 0.60 is reported in the literature for the reliability of single ratings, we can calculate that the ICC would be 0.75 when using the average score of two raters and 0.82 when using the average score of three raters. This is the advantage of using the SB formula.

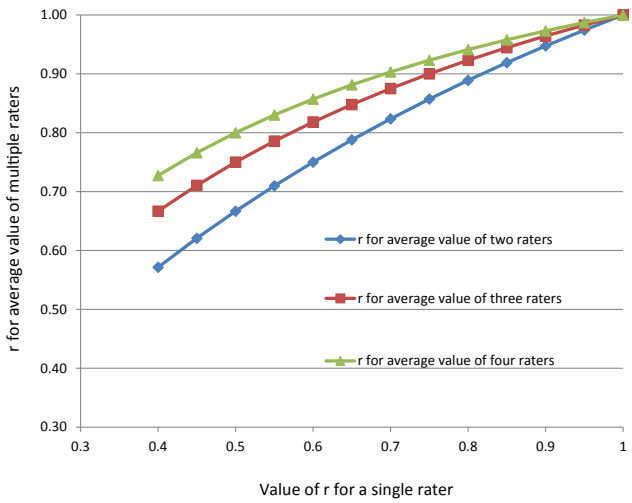

Fig. 1. Correspondence between reliability coefficients for single raters (x-axis) with reliability coefficients for the average score of two, three, or four raters (y-axis).

4.2. Use "Cronbach's alpha if item deleted" to determine the quality of the raters

In multi-item instruments, we often try to improve the reliability in terms of internal consistency by using the option "Cronbach's alpha if item deleted" in SPSS. This test can also be used to select the best raters for a specific task. For example, if we want to select a number of raters for scoring hundreds of images in a large multicenter study, we can select raters that score images as much as possible in a similar way. It is possible to have a test set of images scored by all raters in an interrater reliability study. As we are interested in the performance of these specific raters, we consider the raters as fixed. We select those raters who rate the images in a similar way based on "Cronbach's alpha if item deleted." In the example of SymNose, we calculated Cronbach's alpha and examined with "Cronbach's alpha if item deleted" in which rater performs least consistently, that is, the rater with the highest Cronbach's alpha if item deleted (see Appendix A.5 at www.jclinepi.com for the output). We have never seen this application, but it might be suitable as a pilot study for a large-scale study that includes subjective ratings by multiple raters.

5. Assumptions and remarks

5.1. Assumptions regarding random and fixed factors

The two applications that we described in this paper are based on different assumptions with respect to "random" or "fixed" factors. Using the SB formula to estimate how many raters are required assumes that the extra raters (or the ones that are omitted) can be considered random raters who all perform similarly. However, the application "Cronbach's alpha when item deleted" only makes sense if there is a fixed set of raters to choose from. The aim is to delete the least consistently performing rater(s) to improve reliability.

5.2. Assumptions on the representativeness of raters

In our example, the estimation of the variance components and the calculation of the ICC values were based on the scores of four raters. Quite often in interrater

reliability studies, there are only two raters who are considered to be a random sample of all raters. Appendix A.6 at www.jclinepi.com contains a table with the ICCs based on all possible pairs of raters than can be formed with these four raters. The ICC for a single measurement varies from 0.509 to 0.858, depending on which raters are used. Using the SB formula, the estimated ICC value for the average score of four rates ranges from 0.806 to 0.960.

This means that a design based on only two raters provides a less robust basis for the estimation of reliability than a design with more raters. The usual design of having two raters in a reliability study and calculating the single measure reliability coefficient based on the ratings of only two raters is a poor habit. It is better to estimate in a reliability study the single measure reliability based on ratings of four raters and conclude that the average value of two raters will do than to estimate the single measure reliability on ratings of two raters and then calculate how many raters are needed to get a sufficiently reliable measurement.

5.3. Lessons learned

The two applications we described in this paper provide a deeper insight into reliability and emphasize the similarity between the different forms of reliability. Both applications make use of similarity of concepts on which internal consistency and interrater reliability are based.

The principle of ''Cronbach's alpha if item deleted'' has a limited application. It might be applied in a pilot study to select raters from a fixed set of raters to perform measurements in a large-scale study.

The SB formula has a broader application, and it points to an important lesson for the design and application of reliability studies. With respect to the design, we have to remind the reader that reliability studies aim to estimate the relevant variances between patients, raters, and error. Using more than two raters in a reliability study results in a more robust estimation of the variance components. A sample of two ''random'' raters can hardly be considered to be representative of all raters. Unfortunately, this is the most frequently used design in reliability studies.

Reliability studies often conclude by stating whether the reliability is acceptable or not, without taking further action. Using the SB formula creates the opportunity to estimate how many ratings per person or how many different raters' scores are needed to be averaged to achieve acceptable reliability and then consider whether this is feasible. For example, when a reliability coefficient of 0.65 is found for rating nose symmetry by a single rater, the authors might suggest to always use the average score of two raters in clinical practice. In that case, the reliability rises to 0.79 (see also Fig. 1). When four raters are needed to obtain an acceptable reliability, it is clear that this is logistically impossible in clinical care. By showing these consequences, the SB formula enhances the clinical application of the results of reliability assessments.

We hope that this paper has increased the insight into reliability studies, its assumptions, and applications.

Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.jclinepi.2017.01.013.

References

[1] Mokkink LB, Terwee CB, Patrick D, Alonso J, Stratford P, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol 2010; 63:737–45.
[2] De Vet HCW, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. J Clin Epidemiol 2006;59:1033–9.
[3] Lord FM, Novick MR. Statistical theories of mental test scores. Reading, MA: Addison-Wesley; 1968.
[4] McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychol Methods 1996;1:30–46.
[5] Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951;16:297–334.
[6] Cortina JM. What is coefficient alpha? An examination of theory and applications. J Appl Psychol 1993;78:98–104.
[7] Warrens MJ. Some relationships between Cronbach's alpha and the Spearman- Brown formula. J Classification 2015;32:127–37.
[8] Spearman C. Correlation calculated from faulty data. Br J Psychol 1910;3:271–95.
[9] Brown W. Some experimental results in the correlation of mental abilities. Br J Psychol 1910;3:296–322.
[10] Pigott RW, Pigott BB. Quantifying asymmetry and scar quality of children with repaired cleft lip and palate using Symnose. Cleft Palate Craniofac J 2016;53:298–301.
[11] Mosmuller D, Tan R, Mulder F, Bachour Y, de Vet H, Don Griot P. The use and reliability of SymNose for quantitative measurement of the nose and lip in unilateral cleft lip and palate patients. J Craniomaxillofac Surg 2016;44:1515–21.