A comparison of reliability coefficients for psychometric tests that consist of two parts

Warrens, Matthijs J.

Published in:
Advances in Data Analysis and Classification

DOI:
10.1007/s11634-015-0198-6

Publication date:
2016

Adv Data Anal Classif (2016) 10:71–84
DOI 10.1007/s11634-015-0198-6

REGULAR ARTICLE

A comparison of reliability coefficients for psychometric tests that consist of two parts

Matthijs J. Warrens

Received: 10 October 2013 / Revised: 5 January 2015 / Accepted: 26 January 2015 /
Published online: 8 February 2015

Abstract If a test consists of two parts the Spearman–Brown formula and Flanagan's coefficient (Cronbach's alpha) are standard tools for estimating the reliability. However, the coefficients may be inappropriate if their associated measurement models fail to hold. We study the robustness of reliability estimation in the two-part case to coefficient misspecification. We compare five reliability coefficients and study various conditions on the standard deviations and lengths of the parts. Various conditional upper bounds of the differences between the coefficients are derived. It is shown that the difference between the Spearman–Brown formula and Horst's formula is negligible in many cases. We conclude that all five reliability coefficients can be used if there are only small or moderate differences between the standard deviations and the lengths of the parts.

Keywords Spearman–Brown formula · Cronbach's alpha · Flanagan's coefficient · Angoff–Feldt coefficient · Raju's beta · Horst's formula

Mathematics Subject Classification 62H20 · 62P15 · 91C99

1 Introduction

In psychometrics researchers are concerned with measuring knowledge, abilities and attitudes of persons and individuals. To measure these types of constructs investigators use measurement instruments like tests, exams and questionnaires. In this paper we will refer to any measurement instrument as a test. In test theory, an important concept of a test is its reliability, which indicates how precise a participant's score

M. J. Warrens (✉)
Unit Methodology and Statistics, Institute of Psychology, Leiden University,
P.O. Box 9555, 2300 RB Leiden, The Netherlands
e-mail: warrens@fsw.leidenuniv.nl

is measured. In general, a test is said to be reliable if it produces similar scores for participants under consistent conditions. In reliability estimation a researcher wants to reflect the impact of as many sources of measurement error as possible (Feldt and Brennan 1989). For example, to reflect the day-to-day variation in efficiency of human minds it is an acknowledged principle that a researcher uses at least two interchangeable test forms (parallel-forms approach) or administers the same test twice (test–retest approach). However, because multiple testing is often considered too demanding for the participants, too time-consuming, or too costly, investigators usually do only one test administration. If there is only one test administration researchers may resort to, in the context of classical test theory (Lord and Novick 1968), internal consistency coefficients for estimating the reliability of the test. The most commonly used consistency coefficients are Cronbach's alpha and the Spearman–Brown formula (Cortina 1993; Osburn 2000; Hogan et al. 2000; Feldt and Charter 2003; Grayson 2004; Warrens 2014, 2015).

Internal consistency coefficients estimate reliability by dividing the total test into parts. A test may already consist of multiple parts, for example, a multiple choice part and an essay part. If the test consists of a set of items, the parts can be the individual items or subsets of the items. All reliability coefficients are based on the assumption that the different parts are homogeneous in content (Feldt and Brennan 1989). However, the coefficients are based on different conceptions of how the parts are related. For the coefficients in this paper there are three relevant measurement models, namely, classical parallel, essential tau-equivalence, and congeneric. The models are further discussed in the next section.

In this paper we compare five internal consistency coefficients that can be used if the test is divided into two parts. The coefficients are, the Spearman–Brown formula (Spearman 1910; Brown 1910), Flanagan's coefficient (Rulon 1939), Horst's formula (Horst 1951), the Angoff–Feldt coefficient (Angoff 1953; Feldt 1975), and Raju's beta (Raju 1977). The well-known coefficient Cronbach's alpha reduces to Flanagan's coefficient if we have only two parts. There are several reasons why a test can be divided in only two parts. Sometimes the requirement of content equivalence between parts limits the number of parts to two. Furthermore, in performance and educational settings tests frequently consist of a multiple choice part and an essay part. If previous research has shown that the two parts tend to measure the same construct of interest, it makes sense to leave the two parts intact in reliability estimation. Finally, with two parts we have the simplest reliability formulas and only a few statistics need to be calculated, which may also be a consideration.

The Spearman–Brown formula is based on the classical model, whereas Flanagan's coefficient is based on the essential tau-equivalence approach. The other three coefficients can be used if the more general congeneric model holds. If a test consists of two parts the Spearman–Brown formula and Flanagan's coefficient (Cronbach's alpha) are commonly used, even when their associated measurement models may fail to hold. It appears that researchers are not aware that these coefficients are not universally applicable (Feldt and Charter 2003). Since it is likely that researchers will continue to use the Spearman–Brown formula and Flanagan's coefficient in the near future, it seems useful to study how robust reliability estimation in the two-part case is to coefficient misspecification. We will do this by determining conditions under which

the five coefficients produce (very) similar values, that is, conditions under which the coefficients can be used interchangeably.

Using simulated data Osburn (2000) and Feldt and Charter (2003) showed that the Spearman–Brown formula, Flanagan's coefficient, the Angoff–Feldt coefficient and Raju's beta produce similar values in a variety of situations. In this paper we compare the five coefficients analytically and derive upper bounds of the differences between the coefficients. The upper bounds hold under certain conditions on the standard deviations and lengths of the parts. In the process we formalize several rules of thumb presented in Feldt and Charter (2003). The paper is organized as follows. In the next section we introduce notation, discuss four measurement models, and define the five reliability coefficients. Unconditional and conditional inequalities between the coefficients are presented in Sect. 3. In Sect. 4 we study the pairwise differences between several coefficients and derive upper bounds of the differences and associated conditions. Section 5 contains a discussion.

2 Notation and definitions

In this section we introduce notation and define the five internal consistency coefficients. The coefficients are applicable to tests that consist of two parts, denoted by X_1 and X_2. Each part provides a part score for each participant. If we add the two part scores we obtain the total score $X = X_1 + X_2$. In classical test theory it is assumed that the total score can be decomposed into a true score T and an error term E, $X = T + E$. It is assumed that the error term is unrelated to the true score. The reliability of the total score is defined as the ratio of the true score variance to the total variance:

$$\rho = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2} = \frac{\sigma_T^2}{\sigma_X^2}, \tag{1}$$

where σ_T^2 is the variance of T, σ_E^2 is the variance of E, and σ_X^2 is the total variance of X.

Because the variance of the true score is not known, Eq. (1) cannot be used to estimate the reliability of X. The reliability can be estimated by examining the relationships among the parts. It is assumed that each part is the sum of a true score and an error term, $X_1 = T_1 + E_1$ and $X_2 = T_2 + E_2$. The true score of the total score is the sum of the true scores of the parts, $T = T_1 + T_2$, and the error term of the total score is the sum of the error terms of the parts, $E = E_1 + E_2$. Additional assumptions about T_1, T_2, E_1 and E_2 define different measurement models from classical test theory. Four measurement approaches and their assumptions are presented in Table 1. We will briefly describe the measurement approaches. See, for example, Feldt and Brennan (1989) or Feldt and Charter (2003) for full descriptions of the models.

In the parallel measurement approach it is assumed that each participant has the same true score for both parts, and that the parts have equal error variances. This parallel model is the most restrictive model. If the observed variances of the parts differ extremely then the classical parallel model fails to hold. Several alternative models relax the notion of parallelism. In the tau-equivalent approach the parts may have

Table 1 Assumptions of four measurement models from classical test theory

Measurement model	True scores	Error terms
Parallel	$T_1 = T_2$	$\text{var}(E_1) = \text{var}(E_2)$
	$\text{var}(T_1) = \text{var}(T_2)$	
Tau-equivalence	$T_1 = T_2$	$\text{var}(E_1) \neq \text{var}(E_2)$
	$\text{var}(T_1) = \text{var}(T_2)$	
Essential	$T_1 = T_2 + b_1$	$\text{var}(E_1) \neq \text{var}(E_2)$
Tau-equivalence	$\text{var}(T_1) = \text{var}(T_2)$	
Congeneric	$T_1 = b_2 T_2 + b_3$	$\text{var}(E_1) \neq \text{var}(E_2)$
	$b_2^2 \sigma_T^2 \neq b_3^2 \sigma_T^2$	

where b_1, b_2 and b_3 are real numbers

different error variances. Moreover, if the parts are essentially tau-equivalent it is also allowed that the true scores of a participant on the parts differ by an additive constant. This constant is the same for every participant. If the parts have substantially different lengths it is unrealistic to assume essential tau-equivalence. Finally, the congeneric model is the most general model. In this approach the true score variances of the parts are also allowed to be different.

The remainder of this section is used to introduce five reliability coefficients. Each coefficient is an estimate of the reliability of the total score (1). Let the variance of X_1 and X_2 be denoted by σ_1^2 and σ_2^2, respectively, and let the covariance and correlation between the part scores be denoted by σ_{12} and r, respectively. Between these statistics and the variance of the total score σ_X^2 we have the identity

$$\sigma_X^2 = \sigma_1^2 + \sigma_2^2 + 2\sigma_{12}, \tag{2}$$

and the well-known identity

$$r = \frac{\sigma_{12}}{\sigma_1 \sigma_2}. \tag{3}$$

The Horst and Raju approaches assume that the lengths of the parts are known. This information is often reflected in the number of items (questions, exercises) in each part. Let n_1 and n_2 represent these lengths in an appropriate metric, and let $p_1 = n_1/(n_1 + n_2)$ and $p_2 = n_2/(n_1 + n_2)$. Quantities p_1 and p_2 are the proportions of items in part 1 and 2, respectively.

The oldest and perhaps the most well-known coefficient for two parts is the Spearman–Brown formula (Spearman 1910; Brown 1910) defined as

$$SB = \frac{2r}{1 + r}. \tag{4}$$

In the derivation of this coefficient it is assumed that the two parts are classically parallel. If the variances σ_1^2 and σ_2^2 differ substantially then this model does not hold. The second coefficient is Flanagan's coefficient (Rulon 1939) defined as

$$\alpha = \frac{4\sigma_{12}}{\sigma_X^2}. \tag{5}$$

Coefficient (5) is denoted by α because Cronbach's alpha reduces to this formula if there are only two parts. Coefficient alpha was proposed by Guttman (1945) as lambda3 and later popularized as coefficient alpha by Cronbach (1951). In the derivation of this coefficient the essential tau-equivalence model is assumed, a model that is more flexible than the parallel model. If p_1 and p_2 differ substantially then this model does not hold. In this case it is more appropriate to apply one of the following three coefficients.

Horst's formula, the Angoff–Feldt coefficient, and Raju's beta can be used if it assumed that the most general measurement model, the congeneric model, holds (Feldt and Brennan 1989). Horst's formula is defined as

$$H = \frac{r\sqrt{r^2 + 4p_1p_2(1-r^2)} - r^2}{2p_1p_2(1-r^2)}. \tag{6}$$

Horst (1951) proposed formula (6) as an alternative of the Spearman–Brown formula. It can be used as an estimate of the reliability when the two parts have not necessarily the same length. In version 20 of the software package IBM SPSS Statistics this formula is called the 'Unequal Length Spearman–Brown' (IBM 2011). For $p_1 = p_2 = \frac{1}{2}$ the formula reduces to the Spearman–Brown formula (Horst 1951). Thus, coefficient (4) is a special case of coefficient (6).

The Angoff–Feldt coefficient is defined as

$$AF = \frac{4\sigma_{12}}{\sigma_X^2 - \dfrac{(\sigma_1^2 - \sigma_2^2)^2}{\sigma_X^2}}. \tag{7}$$

It was proposed independently by Angoff (1953) and Feldt (1975). Since coefficient (7) does not depend on p_1 and p_2, it can be used when the lengths of the parts are unknown. If $\sigma_1^2 = \sigma_2^2$ the Angoff–Feldt coefficient reduces to Flanagan's coefficient. Thus, coefficient (5) is a special case of coefficient (7).

Finally, Raju (1977) proposed coefficient beta defined as

$$\beta = \frac{\sigma_{12}}{p_1p_2\sigma_X^2}. \tag{8}$$

Coefficient (8) can be used if the two parts have not necessarily the same length. If we have $p_1 = p_2$ then Raju's beta reduces to Flanagan's coefficient. Furthermore, if we set

$$p_1 = \frac{\sigma_1^2 + \sigma_{12}}{\sigma_X^2} \quad \text{and} \quad p_2 = \frac{\sigma_2^2 + \sigma_{12}}{\sigma_X^2}, \tag{9}$$

then coefficient beta reduces to the Angoff–Feldt coefficient (Feldt and Brennan 1989). Thus, both coefficients (5) and (7) are special cases of coefficient (8). The number of items in each part, n_1 and n_2, may not be good indicators of the relative importance of the parts. In this case the values in (9) can be used instead.

3 Inequalities

In this section we present several inequalities between the five reliability coefficients. It turns out that Flanagan's coefficient is a lower bound of the other four coefficients. Some of the inequalities below have already been demonstrated by other authors. However, in this paper we are also interested in the conditions that specify when the inequalities are equalities. These conditions are often not specified in the literature. The formulations of the inequalities in this section therefore give a more complete picture of how the reliability coefficients are related.

Raju (1977) proved the inequality $\alpha \leq \beta$.

Lemma 1 *$\alpha \leq \beta$ with equality if and only if $p_1 = p_2 = \frac{1}{2}$.*

Proof We have $\alpha \leq \beta \Leftrightarrow 4p_1p_2 \leq 1$. □

The double inequality $\alpha \leq SB \leq AF$ is demonstrated in, for example, Feldt and Charter (2003).

Lemma 2 *$\alpha \leq SB$ with equality if and only if $\sigma_1 = \sigma_2$.*

Proof Using identity (3) we can write the Spearman–Brown formula as

$$SB = \frac{2\sigma_{12}}{\sigma_1\sigma_2 + \sigma_{12}}. \tag{10}$$

Using (10) we have $\alpha \leq SB \Leftrightarrow \sigma_1^2 + \sigma_2^2 \geq 2\sigma_1\sigma_2 \Leftrightarrow (\sigma_1 - \sigma_2)^2 \geq 0$. □

Lemma 3 *$SB \leq AF$ with equality if $r = 1$ or if $\sigma_1 = \sigma_2$.*

Proof Feldt and Charter (2003, p 107) showed that we may write AF as

$$AF = \frac{2r}{1 + r - \dfrac{(1-r)(\sigma_1 - \sigma_2)^2}{\sigma_X^2}}. \tag{11}$$

□

Horst (1951) showed that SB is a special case of H for $p_1 = p_2 = \frac{1}{2}$. Theorem 4 shows that the inequality $SB \leq H$ holds.

Theorem 4 *$SB \leq H$ with equality if and only if $p_1 = p_2 = \frac{1}{2}$.*

Proof We have the identity

$$\begin{aligned} &\frac{1}{2}\left(r - \sqrt{r^2 + 4p_1p_2(1-r^2)}\right)^2 \\ &\quad = r^2 + 2p_1p_2(1-r^2) - r\sqrt{r^2 + 4p_1p_2(1-r^2)}. \end{aligned} \tag{12}$$

Furthermore, the first order partial derivative of H with respect to $p_1 p_2$ is given by

$$\frac{\partial H}{\partial p_1 p_2} = \frac{\dfrac{2p_1 p_2 r\left(1-r^2\right)}{\sqrt{r^2+4p_1 p_2\left(1-r^2\right)}} - r\sqrt{r^2+4p_1 p_2\left(1-r^2\right)} + r^2}{2p_1^2 p_2^2\left(1-r^2\right)}.$$

Multiplying all terms by $\sqrt{r^2+4p_1 p_2\left(1-r^2\right)}$, and using identity (12), we obtain

$$\frac{\partial H}{\partial p_1 p_2} = \frac{-r\left(r-\sqrt{r^2+4p_1 p_2\left(1-r^2\right)}\right)^2}{4p_1^2 p_2^2\left(1-r^2\right)\sqrt{r^2+4p_1 p_2\left(1-r^2\right)}} \leq 0,$$

with equality if and only if $r = 1$ or $r = 0$. Hence, for $r \in (0, 1)$ the function H is strictly decreasing in $p_1 p_2 = p_1(1 - p_1)$, or unimodal in p_1 with a minimum value at $p_1 = p_2 = \frac{1}{2}$ and maximum values close to $p_1 = 1$ and $p_1 = 0$. □

Combining some of the lemmas from this section, we obtain several interesting corollaries. For example, if $p_1 = p_2 = \frac{1}{2}$ we have the double inequality

$$\alpha = \beta \leq SB = H \leq AF.$$

Furthermore, if $\sigma_1 = \sigma_2$ we have the inequality

$$\alpha = SB = AF \leq H, \beta.$$

4 Upper bounds of the differences

In this section we study differences between the five reliability coefficients. In the previous section we presented several inequalities between the coefficients. These results can be used to interpret positive differences between some of the coefficients. For each difference we present several upper bounds and associated conditions.

It follows from Lemma 1 that the difference $\beta - \alpha$ is non-negative. We have the following upper bounds for the difference $\beta - \alpha$.

Lemma 5

$$\beta - \alpha \leq \begin{cases} 0.01 & \text{if } |p_1 - p_2| \leq 0.10; \\ 0.04 & \text{if } |p_1 - p_2| \leq 0.20; \\ 0.09 & \text{if } |p_1 - p_2| \leq 0.30. \end{cases}$$

Proof Using formulas (5) and (8), and the inequality $\beta \leq 1$, we have

$$\beta - \alpha = \beta(1 - 4p_1 p_2) \leq 1 - 4p_1 p_2. \tag{13}$$

The quantity $1-4p_1p_2$ on the right-hand side of (13) is strictly decreasing in $p_1p_2 = p_1(1-p_1)$, or unimodal in p_1 with minimum value zero at $p_1 = p_2 = \frac{1}{2}$ and maximum value unity if $p_1 = 0$ or $p_1 = 1$. For $|p_1 - p_2| = 0.10$ we have $p_1p_2 = 0.2475$ and $4p_1p_2 = 0.99$, for $|p_1 - p_2| = 0.20$ we have $p_1p_2 = 0.24$ and $4p_1p_2 = 0.96$, and for $|p_1 - p_2| = 0.30$ we have $p_1p_2 = 0.2275$ and $4p_1p_2 = 0.91$. This completes the proof. □

Lemma 5 shows that if there are only small differences between the lengths of the parts, Raju's beta and Flanagan's coefficient produce very similar values. If the longest part is one and a half times longer than the shortest part ($|p_1 - p_2| = 0.20$) the difference is at most 0.04. In many applications a difference of this size is of no practical significance.

For Theorems 6 and 7 below it is convenient to work with the ratio

$$c = \frac{\max\{\sigma_1, \sigma_2\}}{\min\{\sigma_1, \sigma_2\}}. \tag{14}$$

The critical values $c = 1.15$ and $c = 1.30$ in Theorems 6 and 7 below are suggested in Feldt and Charter (2003, p 106). It follows from Lemma 2 that the difference $SB - \alpha$ is non-negative. We have the following upper bounds for the difference $SB - \alpha$.

Theorem 6

$$SB - \alpha \leq \begin{cases} 0.0097 & \text{if} \quad c \leq 1.15; \\ 0.0335 & \text{if} \quad c \leq 1.30; \\ 0.0770 & \text{if} \quad c \leq 1.50; \\ 0.0065 & \text{if} \quad c \leq 1.15 \quad \text{and} \quad r \geq 0.50; \\ 0.0110 & \text{if} \quad c \leq 1.30 \quad \text{and} \quad r \geq 0.50; \\ 0.0527 & \text{if} \quad c \leq 1.50 \quad \text{and} \quad r \geq 0.50, \end{cases}$$

where c is defined in (14).

Proof Using formulas (5) and (10) together with equality (2), we have the identity

$$SB - \alpha = SB\left(1 - \frac{2\sigma_1\sigma_2 + 2\sigma_{12}}{\sigma_X^2}\right) = SB\frac{(\sigma_1 - \sigma_2)^2}{\sigma_1^2 + \sigma_2^2 + 2\sigma_{12}}.$$

Since $SB \leq 1$ we have the inequality

$$SB - \alpha \leq \frac{(\sigma_1 - \sigma_2)^2}{\sigma_1^2 + \sigma_2^2 + 2\sigma_{12}}. \tag{15}$$

The right-hand side of (15) is increasing in $|\sigma_1 - \sigma_2|$, or equivalently, increasing in c. Using (14), or $\max\{\sigma_1, \sigma_2\} = c \ \min\{\sigma_1, \sigma_2\}$, in (15) we obtain the inequality

$$SB - \alpha \leq \frac{(1-c)^2 \min\left\{\sigma_1^2, \sigma_2^2\right\}}{\left(1 + c^2\right)\min\left\{\sigma_1^2, \sigma_2^2\right\} + 2\sigma_{12}}. \tag{16}$$

The right-hand side of (16) is decreasing in σ_{12}. Hence, for $\sigma_{12} = 0$ we obtain

$$SB - \alpha \leq \frac{(1-c)^2}{1+c^2}. \tag{17}$$

Using $c = 1.15$, $c = 1.30$ and $c = 1.50$ in (17) we obtain the top three inequalities.

Next, using the identity $r = 0.50$, or equivalently,

$$\sigma_{12} = 0.50\,\sigma_1\sigma_2 = 0.50\,c\ \min\left\{\sigma_1^2, \sigma_2^2\right\} \tag{18}$$

in (15) we obtain

$$SB - \alpha \leq \frac{(1-c)^2}{1+c^2+c}. \tag{19}$$

Since the right-hand side of (16) is decreasing in σ_{12}, we obtain the bottom three inequalities by using $c = 1.15$, $c = 1.30$ and $c = 1.50$ in (19). □

It follows from Lemmas 2 and 3 that the difference $AF - \alpha$ is non-negative. We have the following upper bounds for the difference $AF - \alpha$.

Theorem 7

$$AF - \alpha \leq \begin{cases} 0.0193 & \text{if} \quad c \leq 1.15; \\ 0.0658 & \text{if} \quad c \leq 1.30; \\ 0.1480 & \text{if} \quad c \leq 1.50; \\ 0.0111 & \text{if} \quad c \leq 1.15 \quad \text{and} \quad r \geq 0.50; \\ 0.0384 & \text{if} \quad c \leq 1.30 \quad \text{and} \quad r \geq 0.50; \\ 0.0884 & \text{if} \quad c \leq 1.50 \quad \text{and} \quad r \geq 0.50, \end{cases}$$

where c is defined in (14).

Proof Using (5) and (7) we have the identity

$$AF - \alpha = AF\frac{\left(\sigma_1^2 - \sigma_2^2\right)^2}{\sigma_X^4}.$$

Since $AF \leq 1$ we have the inequality

$$AF - \alpha \leq \frac{\left(\sigma_1^2 - \sigma_2^2\right)^2}{\sigma_X^4}. \tag{20}$$

Using (14), or $\max\{\sigma_1, \sigma_2\} = c\ \min\{\sigma_1, \sigma_2\}$, we have

$$\left(\sigma_1^2 - \sigma_2^2\right)^2 = \left(1 - c^2\right)^2 \min\left\{\sigma_1^4, \sigma_2^4\right\}$$

and

$$\begin{aligned}\sigma_X^4 &= \left(\left(1+c^2\right)\min\left\{\sigma_1^2,\sigma_2^2\right\}+2\sigma_{12}\right)^2\\ &= \left(1+c^2\right)^2\min\left\{\sigma_1^4,\sigma_2^4\right\}+4\sigma_{12}^2+2\left(1+c^2\right)\sigma_{12}\min\left\{\sigma_1^2,\sigma_2^2\right\}.\end{aligned}$$

Hence, (20) can be written as

$$AF-\alpha \leq \frac{\left(1-c^2\right)^2\min\left\{\sigma_1^4,\sigma_2^4\right\}}{\left(1+c^2\right)^2\min\left\{\sigma_1^4,\sigma_2^4\right\}+4\sigma_{12}^2+2\left(1+c^2\right)\sigma_{12}\min\left\{\sigma_1^2,\sigma_2^2\right\}}. \quad (21)$$

The right-hand side of (21) is decreasing in σ_{12}. Hence, for $\sigma_{12}=0$ we obtain

$$AF-\alpha \leq \frac{\left(1-c^2\right)^2}{\left(1+c^2\right)^2}. \quad (22)$$

Using $c=1.15$, $c=1.30$ and $c=1.50$ in (22) we obtain the top three inequalities.

Next, using (18) in (21) we obtain

$$AF-\alpha \leq \frac{\left(1-c^2\right)^2}{\left(1+c^2\right)^2+c^2+c\left(1+c^2\right)}. \quad (23)$$

Since the right-hand side of (21) is decreasing in σ_{12}, we obtain the bottom three inequalities by using $c=1.15$, $c=1.30$ and $c=1.50$ in (23). □

Theorem 7 shows that for small differences between the standard deviations Flanagan's coefficient and the Angoff–Feldt coefficient produce very similar values. If the larger standard deviation is no more than 15 % larger than the smaller ($c \leq 1.15$) then the difference is always less than 0.02. Since the value of the Spearman–Brown formula is between the values of these two coefficients (Lemmas 2 and 3), we may conclude that for $c \leq 1.15$ the difference between the three coefficients is always less than 0.02. In many cases a difference of this size is negligible.

It follows from Theorem 4 that the difference $H-SB$ is non-negative. We have the following upper bounds for the difference $H-SB$.

Theorem 8

$$H-SB \leq \begin{cases} 0.0018 & \text{if } |p_1-p_2| \leq 0.10;\\ 0.0071 & \text{if } |p_1-p_2| \leq 0.20;\\ 0.0162 & \text{if } |p_1-p_2| \leq 0.30;\\ 0.0300 & \text{if } |p_1-p_2| \leq 0.40;\\ 0.0494 & \text{if } |p_1-p_2| \leq 0.50. \end{cases}$$

Proof Using (4) and (6) we have

$$H - SB = \frac{r\sqrt{r^2 + 4p_1p_2\left(1-r^2\right)} - r^2}{2p_1p_2\left(1-r^2\right)} - \frac{2r}{1+r}. \tag{24}$$

In the proof of Theorem 4 it was shown that H is strictly decreasing in $p_1p_2 = p_1(1-p_1)$, or unimodal in p_1 with a minimum value at $p_1 = p_2 = \frac{1}{2}$. Since SB is not a function of p_1p_2, difference (24) is also strictly decreasing in $p_1p_2 = p_1(1-p_1)$, or unimodal in p_1 with minimum value zero at $p_1 = p_2 = \frac{1}{2}$.

The first order partial derivative of (6) with respect to r is given by

$$\frac{\partial H}{\partial r} = \frac{\left(\sqrt{r^2 + 4p_1p_2\left(1-r^2\right)} + \frac{r^2(1-4p_1p_2)}{\sqrt{r^2+4p_1p_2(1-r^2)}} - 2r\right)\left(1-r^2\right)}{2p_1p_2\left(1-r^2\right)^2} + \frac{2r^2\left(\sqrt{r^2 + 4p_1p_2\left(1-r^2\right)} - r\right)}{2p_1p_2\left(1-r^2\right)^2}.$$

Multiplying all terms by $\sqrt{r^2 + 4p_1p_2\left(1-r^2\right)}$ we obtain, using identity (12),

$$\begin{aligned}\frac{\partial H}{\partial r} &= \frac{r^2 + 2p_1p_2\left(1-r^2\right) - r\sqrt{r^2 + 4p_1p_2\left(1-r^2\right)}}{p_1p_2\left(1-r^2\right)^2\sqrt{r^2+4p_1p_2(1-r^2)}} \\ &= \frac{\left(r - \sqrt{r^2+4p_1p_2(1-r^2)}\right)^2}{2p_1p_2(1-r^2)^2\sqrt{r^2+4p_1p_2(1-r^2)}}.\end{aligned} \tag{25}$$

Since partial derivative (25) is strictly positive for $r \in (0, 1)$ and $p_1 \in [0, 1]$, coefficient H is strictly increasing in r for $p_1 \in [0, 1]$. Furthermore, SB is also strictly increasing in r. It turns out that difference (24) is unimodal in r with minimum value zero at $r = 0$ and $r = 1$. (To prove this statement one can show that the second partial derivative of (24) with respect to r is negative. The formula is too long to present here and is therefore omitted). Using (25) the first order partial derivative of difference (24) with respect to r is given by

$$\frac{\partial(H-SB)}{\partial r} = \frac{\left(r - \sqrt{r^2 + 4p_1p_2\left(1-r^2\right)}\right)^2}{2p_1p_2\left(1-r^2\right)^2\sqrt{r^2 + 4p_1p_2\left(1-r^2\right)}} - \frac{2}{(1+r)^2}. \tag{26}$$

To find the value of r for which difference (24) is maximal we need to solve $\partial(H - SB)/\partial r = 0$. However, the maximal value of r depends on the value of p_1p_2. For $|p_1 - p_2| = 0.10$ we have $p_1p_2 = 0.2475$ and $4p_1p_2 = 0.99$, and $\partial(H-SB)/\partial r = 0$ becomes

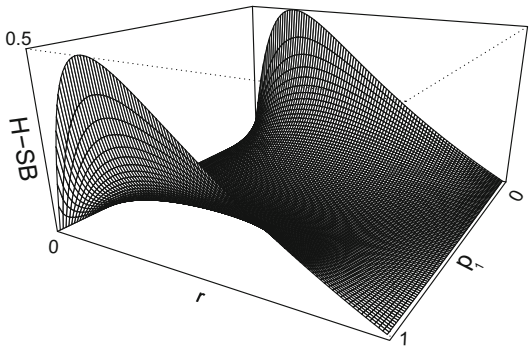

Fig. 1 Plot of the difference $H - SB$ as a function of r and p_1

$$\frac{\left(r - \sqrt{r^2 + 0.99\left(1 - r^2\right)}\right)^2}{0.495\left(1 - r^2\right)^2\sqrt{r^2 + 0.99\left(1 - r^2\right)}} = \frac{2}{(1 + r)^2}. \tag{27}$$

The solution in the range $[0, 1]$ of (27) is $r \approx 0.41335$. Using $p_1 p_2 = 0.2475$ and $r = 0.41335$ in (24) we obtain $H - SB = 0.00172$. Since difference (24) is unimodal in $p_1 p_2$ it follows that $H - SB \leq 0.0018$ if $|p_1 - p_2| \leq 0.10$. The other conditional inequalities are obtained from using similar arguments. □

Theorem 8 shows that even with substantial differences between the lengths of the parts the Spearman–Brown formula and Horst's formula produce very similar values. When the longest part is one and a half times longer than the shortest part ($|p_1 - p_2| = 0.20$) the difference is less than 0.01, a size that is negligible. Even when the longest part is three times longer than the shortest part ($|p_1 - p_2| = 0.50$) the difference is less than 0.05.

Figure 1 presents a 3D plot of difference (24) as a function of r and p_1. The figure visualizes some of the ideas in Theorem 8. Figure 1 shows that the difference $H - SB$ is small for moderate values of p_1, that is, small values of the difference $|p_1 - p_2|$. Furthermore, these small differences do not depend on the value of r.

5 Discussion

In this paper we compared five reliability coefficients for tests that consist of two parts. The coefficients are the Spearman–Brown formula, Flanagan's coefficient (a special case of Cronbach's alpha), Horst's formula, the Angoff–Feldt coefficient, and Raju's beta. We first presented inequalities between the reliability coefficients. The inequalities were then used to formulate positive differences between the coefficients. Using analytical techniques we then derived several upper bounds of the differences between the coefficients. The upper bounds hold under certain conditions.

Criteria for qualifying the values of the differences between the coefficients depend on the context of the reliability estimation. In this paper we use the following criteria. A

difference of at most 0.04 is considered to be of no practical significance. Furthermore, a value that is smaller than or equal to 0.02 is considered to be negligible. A difference between coefficients is substantial if its size is bigger than or equal to 0.10. These criteria are of course arbitrary. The reader may interpret the results using other critical values.

An interesting relationship was found between the values of the Spearman–Brown formula and Horst's formula. Theorem 8 shows that even with substantial differences between the lengths of the parts the formulas produce very similar values. When the longest part is one and a half times longer than the shortest part the difference is less than 0.01. A difference of this size is negligible. But even if the longest part is three times longer than the shortest part the difference between the coefficients is always less than 0.05. The Spearman–Brown formula is based on the classical parallel model, and this model fails to hold with substantial differences between the lengths of the parts. Horst (1951) proposed his formula for the case that the parts have different lengths. However, in many real-life situations the difference will be negligible, although Horst's formula will always produce a (slightly) higher value.

Lemmas 2 and 3 together with Theorem 7 show that for small differences between the standard deviations the Spearman–Brown formula, Flanagan's coefficient and the Angoff–Feldt coefficient produce very similar values. If the larger standard deviation is no more than 15 % larger than the smaller then the difference is always less than 0.02. In this case all three coefficients can be used, which confirms and formalizes a rule of thumb presented in Feldt and Charter (2003). Theorems 6, 7 and 8 together with Lemma 5 show that for small and moderate differences between the lengths and standard deviations of the parts all five coefficients produce very similar values. If the larger standard deviation is no more than 15 % larger than the smaller, and if the difference between the lengths (in proportions) is at most 0.10, then the differences between the values is less than 0.02. In this case all five coefficients can be used. These results partly explains why the coefficients produce very similar values for the simulated data in Osburn (2000).

If the larger standard deviation is no more than 30 % larger than the smaller, and if the difference between the lengths is at most 0.20, then the differences between the values is less than 0.07. If we exclude the Angoff–Feldt coefficient, the differences between the values of the other four coefficients is less than 0.041. If, in addition, the correlation between the parts is at least 0.50, then the differences between all five coefficients is less than 0.04. In this case application of any coefficient will probably lead to the same conclusion. Finally, even if the differences between the standard deviations and lengths are relatively large, the maximum difference between the coefficients is less than 0.10. More precisely, if the larger standard deviation is no more than 50 % larger than the smaller, if the difference between the lengths is at most 0.30, and if the correlation between the parts is at least 0.50, then the differences between the values of the coefficients is at most 0.09.

Since the five coefficients produce very similar values for small and moderate differences between the standard deviations and the lengths of the two parts, we conclude that reliability estimation in the two-part case tends to be robust to coefficient misspecification. If the difference between the standard deviations and the lengths are large the values of the reliability coefficients diverge. In this case both the classical

parallel model and the essential-tau equivalent model fail to hold, and application of the Spearman–Brown formula and Flanagan's coefficient (Cronbach's alpha) is not appropriate. Which coefficient should be used in this case, Horst's formula, the Angoff–Feldt coefficient, or Raju's beta, appears to be an open problem, and is thus a topic for further investigation.

References

Angoff WH (1953) Test reliability and effective test length. Psychometrika 18:1–14
Brown W (1910) Some experimental results in the correlation of mental abilities. Br J Psychol 3:296–322
Cortina JM (1993) What is coefficient alpha? An examination of theory and applications. J Appl Psychol 78:98–104
Cronbach LJ (1951) Coefficient alpha and the internal structure of tests. Psychometrika 16:297–334
Feldt LS (1975) Estimation of reliability of a test divided into two parts of unequal length. Psychometrika 40:557–561
Feldt LS, Brennan RL (1989) Reliability. In: Linn RL (ed) Educational measurement, 3rd edn. Macmillan, New York, pp 105–146
Feldt LS, Charter RA (2003) Estimating the reliability of a test split into two parts of equal or unequal length. Psychol Methods 8:102–109
Grayson D (2004) Some myths and legends in quantitative psychology. Underst Stat 3:101–134
Guttman L (1945) A basis for analyzing test–retest reliability. Psychometrika 10:255–282
Hogan TP, Benjamin A, Brezinski KL (2000) Reliability methods: a note on the frequency of use of various types. Educ Psychol Meas 60:523–561
Horst P (1951) Estimating the total test reliability from parts of unequal length. Educ Psychol Meas 11:368–371
IBM (2011) IBM SPSS Statistics 20 Algorithms Manual, p 1069
Lord FM, Novick MR (1968) Statistical theories of mental test scores. Addison-Wesley, Reading
Osburn HG (2000) Coefficient alpha and related internal consistency reliability coefficients. Psychol Methods 5:343–355
Raju NS (1977) A generalization of coefficient alpha. Psychometrika 42:549–565
Rulon PJ (1939) A simplified procedure for determining the reliability of a test by split-halves. Harv Educ Rev 9:99–103
Spearman C (1910) Correlation calculated from faulty data. Br J Psychol 3:271–295
Warrens MJ (2014) On Cronbach's alpha as the mean of all possible k-split alphas. Adv Stat 2014
Warrens MJ (2015) Some relationships between Cronbach's alpha and the Spearman–Brown formula. J Classif (in press)