Article details

Goeman J.J. & Jong N.H. de (2018), How Well Does the Sum Score Summarize the Test? Summability as a Measure of Internal Consistency, *Educational Measurement: Issues and Practice* 37(2): 54-63.
Doi: 10.1111/emip.12181

Educational Measurement
ISSUES AND PRACTICE

NCME national council on measurement in education

Educational Measurement: Issues and Practice
Summer 2018, Vol. 37, No. 2, pp. 54–63

How Well Does the Sum Score Summarize the Test? Summability as a Measure of Internal Consistency

J. J. Goeman and N. H. De Jong, *Leiden University*

Many researchers use Cronbach's alpha to demonstrate internal consistency, even though it has been shown numerous times that Cronbach's alpha is not suitable for this. Because the intention of questionnaire and test constructers is to summarize the test by its overall sum score, we advocate summability, which we define as the proportion of total test variation that is explained by the sum score. This measure is closely related to Loevinger's H. The mathematical derivation of summability as a measure of explained variation is given for both scale and dichotomously scored items. Using computer simulations, we show that summability performs adequately and we apply it to an existing productive vocabulary test. An open-source tool to easily calculate summability is provided online (https://sites.google.com/view/summability).

Keywords: Cronbach's alpha, internal consistency, Loevinger's H, validity, unidimensionality

Introduction

In many fields, such as education or second language acquisition, tests of various kinds are constructed to measure (language) abilities. In sociological and behavioral research, researchers use questionnaires to measure beliefs, attitudes, personality attributes, and intentions. In the field of experimental psychology, whenever individual differences are investigated, likewise, tests and questionnaires are constructed and used to gauge (cognitive) abilities and skills. All these tests and questionnaires are usually constructed with the explicit intention to summarize the results by the sum score of the items, and for questionnaires, the sum score divided by the number of items. Items for a test (in what follows, we will use the label *test* to refer to both tests and questionnaires) are purposefully selected for their intended contribution to this sum score. The intended summarization by the sum score is virtually universal in test construction and it reflects the assumption that all items should measure the construct and that all items are separate mini-tests of this construct. No single item can measure the construct exclusively: items are noisy in the sense that they may measure other unrelated constructs as well. By taking a sum score of many items, the noise and the contributions of other constructs hopefully cancel out, making the sum score a less noisy test than the mini-tests it was constructed from. This hope by the researcher will only be borne out if the noise in the items is at random. In other words, the unintended but unavoidable additional constructs measured by each item should be as different as possible for each of the items.

J. J. Goeman, Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, The Netherlands; j.j.goeman@lumc.nl. N. H. De Jong, Leiden University Center for Linguistics and Leiden University Graduate School of Teaching, Leiden University, 2311 BV Leiden, The Netherlands; n.h.de.jong@hum.leidenuniv.nl.

Summability, Reliability, Unidimensionality, Internal Consistency, and Homogeneity

If we see each item as a mini-test (or mini-questionnaire), we want to retain as much of the variation measured by these mini-tests as possible. Since the intention is to summarize, for each individual, the whole test by his or her sum score, we want that sum score to capture as much of the information contained in the items as possible. The percentage of variation measured by the test that is retained by the sum score is the crucial quantity here. This, we call the *summability* of the test. It explicitly connects the validity of the individual items to the validity of the test as a whole. To our knowledge, the fraction of variation retained by the sum score has not been explicitly considered before, although, of course, it relates closely to other concepts that are prerequisites of validity, such as reliability, unidimensionality, internal consistency, and homogeneity. Davenport, Davison, Liou, and Love (2015) distinguished reliability, unidimensionality, and internal consistency, and discussed the mathematical relationships between these concepts.

Reliability is, like summability, a concept that is directly related to the sum score. It is defined as correlation between the sum scores obtained for two separate administrations of the test in the same population (for an in-depth discussion of reliability, see Haertel, 2006). Reliability measures how reproducible the variability in the sum score is measured, but it does not relate the variability of the sum score to the total variability of the test. Since sum scores of longer tests are less noisy, reliability is by definition related to test length, as has been noted before (e.g., Cortina, 1993; Green, Lissitz, & Mulaik, 1977; Nunnally, 1978, pp. 227–228). A long test can be very reliable even if it is multidimensional and measures a number of separate attributes. Such a test may be reliable but is not summable, because much information is discarded when the test is reduced to a sum score.

Hattie (1985), in his authoritative review, considered many quantitative measures of unidimensionality. Tests can be more or less unidimensional, and we can define unidimensionality as the extent to which the variability of the test is captured by a single dimension. Defined in this way, unidimensionality is closely related to summability, in the sense that a test that is highly summable is also highly unidimensional. The converse is not necessarily true. For example, consider a test in which some of the items have been inversely coded. This has no effect on unidimensionality since there will still be a strong principal factor in the data. Summability, however, will decrease when there are miscoded items because the principal factor will be less well represented with the sum score. In the extremum, with half of the items miscoded, summability will be zero. Summability therefore is a special form of unidimensionality, namely, unidimensionality in the direction of the sum score.

Internal consistency, loosely speaking, would be the degree to which all items measure the same thing. Unlike for unidimensionality, there has been relatively little effort to define internal consistency in a precise way, and it has thus far mostly been used as an intuitively appealing but loosely defined concept. Only Loevinger (1948) offers a mathematical operationalization of the concept, which she calls homogeneity. In Loevinger's view, tests are homogeneous if subjects with similar sum scores also have similar answer patterns. Cronbach (1951) briefly discusses internal consistency, which he relates to the average correlation between items. These two views are conceptually similar, as similar answer patterns between subjects result in high correlations between items.

We present a complementary view on homogeneity/internal consistency to that of Loevinger and Cronbach, for which we use the term summability. Summability is the extent to which the information measured by the test is captured by the sum score. A highly homogeneous test is summable because if all answer patterns are alike, this answer pattern is well described by the sum score. A highly internally consistent test is summable, because if all items correlate well with each other then all items correlate well with the sum score.

Measuring Summability and Related Concepts

Although the concepts of reliability, unidimensionality, and homogeneity/internal consistency are related, or even may amount to the same thing in some (ideal) cases (Davenport et al., 2015), they are, in fact, distinct (Sijtsma, 2015). In practice, investigators and test administrators are in need of a measure of quality of the sum score as a summary of a test. Quantities that are frequently used generally measure unidimensionality or reliability, rather than summability (or homogeneity/internal consistency).

Cronbach's alpha remains popular as a measure of the quality of the sum score. It is informally regarded by many researchers as a measure of internal consistency and described as such in handbooks (e.g., Bland & Altman, 1997; Larson-Hall, 2010, pp. 170–175; Pallant, 2007, p. 6). It is well known among methodologists, however, that Cronbach's alpha is not a measure of internal consistency but of reliability (and not a very good one at that; see, e.g., Sijtsma, 2009). In fact, several properties of Cronbach's alpha are at odds with its use as a purported measure of internal consistency. In particular, it is known that any desired level of Cronbach's alpha may always be achieved by simply increasing the number of items in a test (Cortina, 1993; Green et al., 1977; Nunnally, 1978, pp. 227–228). This is natural for a measure of reliability, since longer tests are generally more reliable. Indeed, Spearman (1910) and Brown (1910) argue that any level of reliability may be achieved by increasing test length unless the items are uncorrelated. For example, a test of two independent blocks of five items, correlated only .3 within the block, has $\alpha = .61$; the same test with 30 items per block has $\alpha = .91$. The same property holds for other measures of reliability, for example, intraclass correlation coefficients (Shrout & Fleiss, 1979) and generalizability coefficients (Brennan, 2001, p. 13). As measures of reliability, they intentionally increase with test length. Such a property, however, is clearly undesirable for a measure of consistency, since longer tests are not necessarily more consistent.

Unidimensionality can be measured using factor analysis, which models the items as noisy measurements of linear combinations of one or more latent factors. In this model, a test is unidimensional if a single factor is dominant, explaining a large proportion of the variability of the test items. Although measures of unidimensionality, such as omega hierarchical (Zinbarg, Revelle, Yovel, & Li, 2005), can be derived from a factor model (Hattie, 1985), such measures do not have an immediate relationship to the quality of the sum score. The dominant factor is estimated by a linear combination of the item scores that may involve very different weights between items, and even negative weights. Consequently, the dominant factor may, in theory, be only moderately correlated with the sum score. Separate analyses are necessary to check whether any unidimensionality found is indeed sufficiently related to the sum score. Rasch models make the assumption of unidimensionality at the outset of the modeling process, assuming a single underlying latent score. The type of unidimensionality assumed by Rasch models is also immediately related to the sum score, as it has been shown that the sum score is the sufficient statistic for the latent subject ability in the Rasch model (Andrich, 1988, p. 38). However, by assuming unidimensionality as the cornerstone of the model, it becomes impossible to assess the assumption within the model, and also to quantify the extent of unidimensionality. Goodness-of-fit tests for the Rasch model are sometimes used, but these can only quantify the evidence against unidimensionality, not its extent. Goodness-of-fit tests also have the well-known drawbacks that they have too little power for small sample size and excessive power for large sample size.

Measures of internal consistency are less well known than measures of reliability or unidimensionality. The average correlation coefficient was proposed for this purpose by Cronbach (1951) and recently discussed by Davenport et al. (2015). Loevinger's H (1948) is the number of Guttman's (1950) errors divided by the maximal possible number of Guttman errors, where Guttman errors are deviations from the ideal triangular Guttman pattern that would arise if all subjects would have the same answer pattern. Loevinger's H is currently used only in the context of Mokken's (1971) scaling, although it does not require the assumption of a Mokken model.

Any measure of homogeneity/internal consistency must not depend on test length or on the number of subjects performing the test. Moreover, it should be invariant to the difficulty of typical items or to the variation in item difficulty. Using the concept of summability, we will propose such a measure. It takes into account the intention of summarizing the test by the sum score, and gives a single value to assess the amount of

information from the items that is retained by this summary. The new measure unifies Loevinger's and Cronbach's views on homogeneity/internal consistency. Summability is equal to the average correlation coefficient between items in a test if items are interval scale and all item variances are equal, thus generalizing Cronbach's (1951) measure. For dichotomous items, our measure will be a generalization of Loevinger's H (1948). We formulate a more stable variant of that measure using tetrachoric correlations.

Outline

In what follows, we will formalize *summability* mathematically. Its characteristics will then be shown in a number of simulations. We also compare summability to Cronbach's alpha, a measure that is currently used in many fields to decide whether one can summarize a test by its sum score. We then proceed with a real-life application, using the web application of summability that we created using Shiny (Chang, Cheng, Allaire, Xie, & McPherson, 2016). We calculate summability of a productive vocabulary test and show how summability-if-item-deleted can be used to judge whether removal of specific items would lead to a test that is more homogeneous.

Summability

Mathematical Derivation

We define the summability of a test as the extent to which the intrasubject variability that is measured by the items is captured by the sum score. Since we view each item as a mini-test, we want as much as possible of the information in these items to be retained. As calculation of the sum score is a form of data reduction, it is of interest to know how much information is discarded by this operation. Ideally, if all mini-tests are perfect, no information on intraindividual variation is lost by summing. Variability can be lost, however, whenever items are noisy, if items are mislabeled, or when the test lacks unidimensionality.

We propose to calculate summability as the percentage of the total variance in the test that is retained (or explained) by the sum score. It measures a degree of agreement of the items. High summability implies high concordance between items, so that subjects that score high on one item tend to score high on other items as well. Low summability, vice versa, implies low concordance, meaning that different subjects score high on different items. This property links summability to internal consistency. As an explained variation measure, summability will be invariant to the number of items or to variation in their difficulty. It will also be unaffected by the number of subjects, although using too few subjects, of course, will lead to unstable estimates.

Consider first a test with interval scale scores. The variance of the sum score of such a test is given by the sum of all item variances and covariances:

$$c = \sum_i \sigma_i^2 + 2\sum_{i<j} \sigma_{ij},$$

where σ_i^2 is the variance of item i, σ_{ij} is the covariance of items i and j, and the sums are over all (combinations of) items of the test. The variance c is maximally equal to

$$m = \left(\sum_i \sigma_i\right)^2,$$

which happens when correlations between test items are all equal to 1. The variance m can be seen as the total (potential) amount of variance in the test, and the variance c is the amount of that variance used by the sum score. The unadjusted summability we define as the ratio

$$s_u = c/m,$$

the fraction of the potential test variance that is captured by the sum score.

As a ratio of variances, the unadjusted summability has an interpretation similar to that of an R^2 in a regression model: of the variation m potentially present in the test, a fraction s_u is explained by the sum score, while an unexplained fraction $1 - s_u$ remains. The value of s_u is always between 0 and 1, because $0 \leq c \leq m$.

However, the ratio of c and m is analogous to an unadjusted R^2. Like that quantity, it is always positive, even when the items are completely uncorrelated, because c and m are always positive. Particularly for tests with few items the unadjusted summability can be quite high, and for a test of a single item, the unadjusted summability is always equal to 1. We can reduce summability to a quantity that is 0 in expectation if all items are uncorrelated, in analogy to an adjusted R^2 in regression. To do so, we must subtract the part from c that is positive even if all items are uncorrelated. This is given by

$$v = \sum_i \sigma_i^2.$$

Subtracting the same quantity from m, we obtain the adjusted summability, or briefly just summability, given by

$$s = \frac{c - v}{m - v}.$$

The adjusted summability is still equal to 1 if all items are maximally correlated. Unlike the unadjusted summability, it is not strictly positive but can take negative values if many negative correlations occur between items.

The adjusted summability can be alternatively written as a weighted average of the pairwise correlations between the test items. Writing $\sigma_{ij} = \rho_{ij}\,\sigma_i\sigma_j$, where ρ_{ij} is the Pearson correlation between items i and j, and σ_i is the standard deviation of item i, then we obtain

$$s = \frac{\sum_{i<j} \rho_{ij}\sigma_i\sigma_j}{\sum_{i<j} \sigma_i\sigma_j}.$$

That is, s is a weighted average of the correlations ρ_{ij} for all pairs of items i and j, where the weights are given by $\sigma_i\sigma_j$, the product of the standard deviations of the items. In particular, for an imaginary test of just two items, s reduces to the correlation coefficient. If all item variances are equal, s is just the average correlation

$$\bar{r} = \frac{\sum_{i<j} \rho_{ij}}{\frac{1}{2}(K^2 - K)},$$

which was already suggested as a measure of internal consistency (Cronbach, 1951; Davenport et al., 2015).

The expression of summability as a weighted average correlation makes it easy to extend the concept to variables

measured on nominal or ordinal scales by replacing the Pearson correlation by a correlation that is more appropriate for the measurement scale at hand. For dichotomous variables, we suggest using the tetrachoric correlation. Unlike some other correlation measures for dichotomous variables, this correlation is not bounded away from 1, giving a correlation of 1 whenever all subjects that score 1 on one of the questions also score 1 on the other question, regardless of the relative difficulty of the two items. With tetrachoric correlation, we therefore can obtain a summability of 1, which happens if and only if the test matrix has the following structure: ordering the items by difficulty and the subjects by their sum score will result in the ideal "triangular" test score matrix, a Guttman pattern (Guttman, 1950), in which every subject that scores a difficult item correctly will also score every easier item correctly. The more the test matrix will deviate from this ideal triangular form, i.e., the more the groups that score the different items correctly differ in composition, the more the summability will decrease.

Summability is closely related to the homogeneity measure H by Loevinger (1948). This is also defined as a ratio of $c - v$ and the maximal value that $c - v$ may take. Loevinger only considered the case of dichotomous variables, for which she used Pearson correlation instead of tetrachoric. Pearson correlation is bounded away from 1 when the item difficulties are different, and Loevinger corrects for this when calculating H. Loevinger's H therefore can also be written as a weighted average correlation coefficient, but with a particular choice of correlation coefficient: the Pearson correlation divided by the maximal possible value of the Pearson correlation given the item difficulties (Warrens, 2008). In the simulation comparing Loevinger's H and summability below, we shall see that summability based on tetrachoric correlations has a more stable behavior.

An important practical question is what constitutes a high enough summability for a test. This is a difficult question to answer, and only practical experience will allow formulation of rules of thumb, which may well differ between fields. The interpretation of summability as an R^2 will help to get a feeling for an acceptable magnitude of s. A summability of .2, for example, indicates that a major proportion of the intrasubject variability in the test is discarded by the sum score, which suggests too great heterogeneity between test items. Moderate summability around .3 or .4 is weak but might still be acceptable for constructs known to be difficult to measure, while summabilities in the order of magnitude of .5 to .7 should probably be considered good to very good in many fields, suggesting that the majority of the information in the items is retained in the final score. Similar cutoff values were suggested for Loevinger's H by Mokken (1971, p. 185).

An idea of the order of magnitude can also be obtained from a comparison with Cronbach's alpha, to which summability is related. We will look at the simplest case of standardized interval variables only. In this case, Cronbach's alpha (standardized) relates to summability as

$$\alpha = \frac{K s}{1 + (K - 1) s},$$

where K is the number of items in the test. We see that summability is exactly 1 whenever Cronbach's alpha is exactly 1, but that otherwise $s < \alpha$ whenever s is positive. Cronbach's alpha is especially bigger than summability if K is large and s is not too small. A test of 20 items with summability of .2 has an alpha of .83, which increases to .93 for a summability of .4. With 40 items, a summability of .2 is already sufficient to obtain an alpha of .91, and summability .4 increases this to .96.

Summability as a Summary Measure

Summability summarizes a large amount of information into a single number. Necessarily, some information is lost in the process. Consider, for example, two tests that sum six standardized interval scale items. One of the tests has a correlation matrix between the items given by A and the second by B:

$$A = \begin{pmatrix} 1 & .8 & .8 & 0 & 0 & 0 \\ .8 & 1 & .8 & 0 & 0 & 0 \\ .8 & .8 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & .8 & .8 \\ 0 & 0 & 0 & .8 & 1 & .8 \\ 0 & 0 & 0 & .8 & .8 & 1 \end{pmatrix};$$

$$B = \begin{pmatrix} 1 & .32 & .32 & .32 & .32 & .32 \\ .32 & 1 & .32 & .32 & .32 & .32 \\ .32 & .32 & 1 & .32 & .32 & .32 \\ .32 & .32 & .32 & 1 & .32 & .32 \\ .32 & .32 & .32 & .32 & 1 & .32 \\ .32 & .32 & .32 & .32 & .32 & 1 \end{pmatrix}.$$

These two tests have the same low summability of .32. In the first case, summability is low because the test is composed of two independent constructs, each of which measured with high internal consistency (summability within the construct is .8). The second test has low summability because the items measuring the single construct have a correlation of only .32, indicating that the items in the test are noisy. We see that a low summability can be due to diverse reasons, and the reasons for a low summability of a test should be explored by using other methods, e.g., factor analysis or measures for unidimensionality. A high summability, on the other hand, can only be due to strong internal consistency of the test. It is impossible to make an example such as the one above, in which two halves of the test measure independent constructs, with a summability greater than .5. Note that the two tests also have the same Cronbach's alpha of .74.

Simulations

To illustrate the basic properties of summability, we simulated data from a simple Rasch model. In this simulation, we first randomly generated subject abilities for N independent subjects from a normal distribution with mean μ_{subject} and standard deviation τ_{subject} and item difficulties for K independent items from a normal distribution with mean μ_{item} and standard deviation τ_{item}. Next, we calculated the odds for each subject to score correctly on each item according to the Rasch model, which models the probability of a correct score for subject i with ability a_i on item j with difficulty d_j as

$$p_{ij} = \frac{\exp(a_i - d_j)}{1 + \exp(a_i - d_j)},$$

and used these probabilities to simulate item scores for each subject.

In this Rasch model, which is unidimensional by definition, summability should depend only on variation in subject

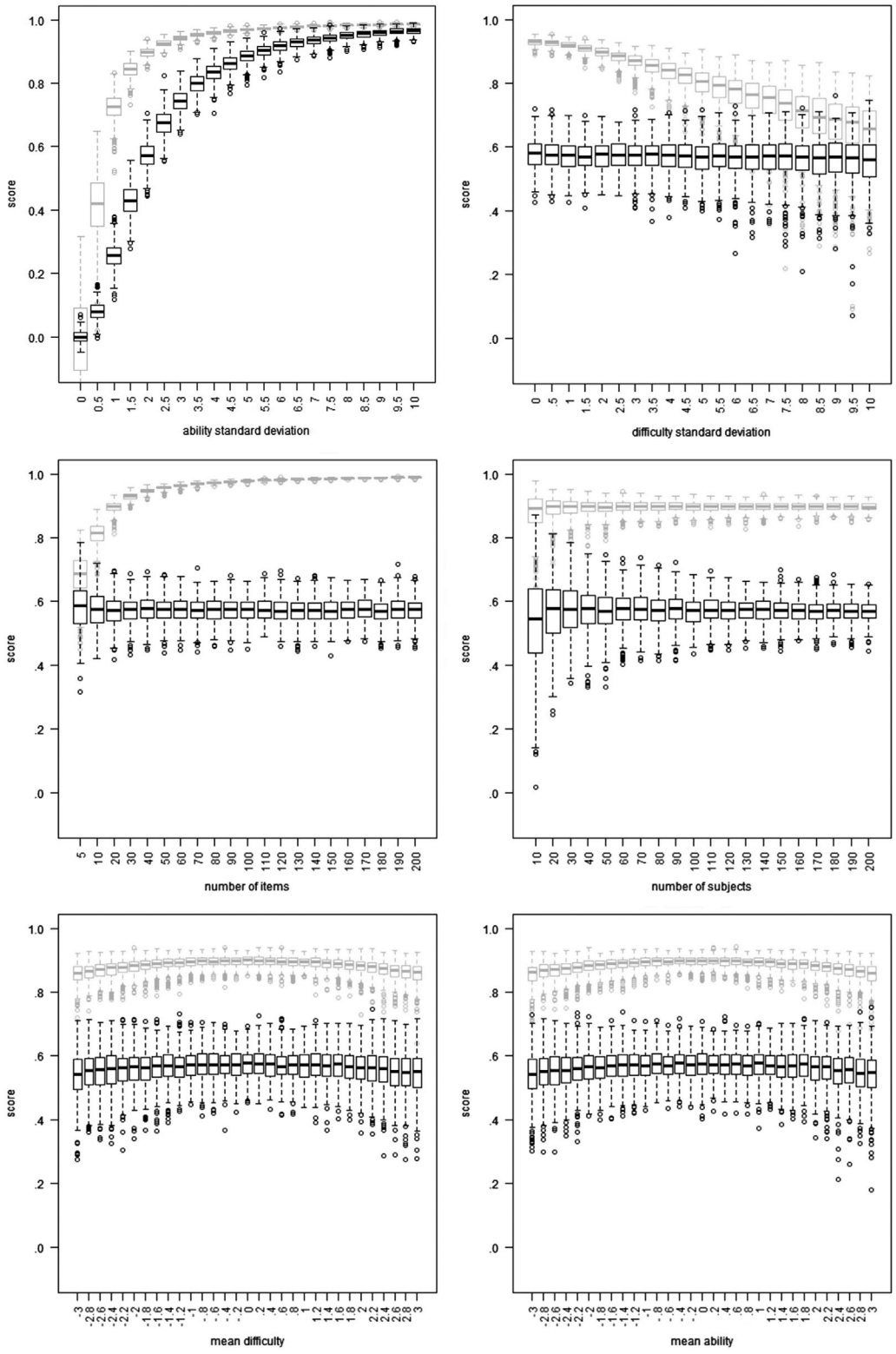

FIGURE 1. Summability (black) and Cronbach's alpha (gray) as a function of various parameters in a simulated Rasch model. All boxplots are based on 500 computer simulations.

ability. If all subjects have equal ability, the test result is just random noise and summability should be zero; the larger the differences in ability, the more signal there is for the test to measure relative to the noise, and the higher summability should be. On the other hand, summability should not depend on any of the other parameters in the model, such as sample size, number of items, average difficulty or ability, or variation in item difficulty.

To test this, we varied the parameters N, K, μ_{subject}, μ_{item}, τ_{subject}, and τ_{item} one at a time from the starting point $N = 100$, $K = 20$, $\mu_{\text{subject}} = 0$, $\mu_{\text{item}} = 0$, $\tau_{\text{subject}} = 2$, and $\tau_{\text{item}} = 2$. The results are given in Figure 1.

We see, as desired, that summability is most dependent on variation in subject ability (shown in the top left panel of Figure 1), which drives the sum score in the Rasch model. At zero standard deviation, summability is close to 0 as there is nothing for the test to measure. With increasing variation in true ability, summability increases toward 1 as test results slowly converge toward an ideal triangular Guttman pattern. Cronbach's alpha has the same behavior but at a much higher level, showing high variability in the low range.

With respect to all other parameters of the model, summability shows remarkable stability. The average level of summability is stable with respect to average item difficulty, average ability, variation in item difficulty, number of items, and number of subjects. More variability in summability arises, naturally, for small numbers of subjects, as well as for tests with few items and for tests with small and large mean difficulty and ability, and large variation in item difficulty, when there are many items for which all subjects score (almost) identically. The stability of summability contrasts starkly with the lack of stability of Cronbach's alpha, which increases sharply to 1 as the number of items increases, and has strong downward bias for tests with extremely difficult or easy items (cf. very able or weak subjects) and especially for tests with a large variability in item difficulty.

In a second simulation, we ran the same analyses, but now on data generated by two independent Rasch models, each on half of the items, so that half of the test measured one construct, while the other half measured a completely independent one. This gives rise to a block-correlation structure. Figure 2 shows the results of this simulation. Here, the major difference is that summability is much lower (about half the value it had in the first simulation), and that it never increases to 1 even if between-subject variability in ability increases, but stays below .5. Cronbach's alpha, in contrast, does increase to 1 in the same situation, seemingly suggesting great internal consistency for the test. The (in)stability in the other parameters is quite the same as in the previous simulation, although summability in this case seems to show a slight downward bias for extremely easy or difficult tests.

Finally, we compared summability with Loevinger's H on the basis of the same Rasch models as in the third simulation. Summability has more stable behavior overall, but especially with respect to variation in item difficulty, as seen in Figure 3 (compare with upper right graph in Figure 1). Unlike summability, Loevinger's H is low when there is little variation in item difficulty and high if there is much variation.

Application

To illustrate the use of summability, we calculated summability for a productive vocabulary test. For the "What Is Speaking Proficiency?" project at the University of Amsterdam, several knowledge and skill tests were administered, in addition to eight speaking tasks. The rationale for this study was to investigate which linguistic knowledge and skills contributed to overall speaking proficiency, and whether the separately measured linguistic knowledge and skills would have different relative weights for the lower proficient and higher proficient speakers (De Jong, Steinel, Florijn, Schoonen, & Hulstijn, 2012; Hulstijn, Schoonen, de Jong, Steinel, & Florijn, 2012). For both groups, the linguistic knowledge and skills could explain a large chunk of the variance in overall speaking proficiency. The best predictor of speaking proficiency turned out to be a measure of vocabulary knowledge. In what follows, we will first describe the participants and materials of the vocabulary task as reported in De Jong et al. (2012) before we report the summability of this test.

Participants. Data were collected from 207 adult second language (L2) learners of Dutch, and 59 native (L1) speakers of Dutch. Most of the learners were taking Dutch courses at intermediate or advanced levels to prepare for enrollment at the University of Amsterdam. The ages ranged from 20 to 56 ($M = 29$; $SD = 6$); 72% were female and 28% were male. The L1 background of the L2 learners was diverse: 46 different L1s were reported, with German ($n = 23$), English ($n = 18$), Spanish ($n = 16$), French ($n = 15$), Polish ($n = 11$), and Russian ($n = 10$) as most frequently reported L1s. Participants' length of residence in the Netherlands ranged from 10 months to 20 years. Most L1 speakers were enrolled at (the same) university, their age ranged from 18 to 45 ($M = 25$; $SD = 6$), 63% were female and 37% were male.

Materials. A two-part paper-and-pencil task was administered. Part 1 (90 items) elicited knowledge of single words, and part 2 (26 items) elicited knowledge of multiword units. In this article, we will only report on the first part of the vocabulary test. For this part, nine words were selected from each frequency band of 1,000 words between words ranked 1–10,000 according to the Corpus Gesproken Nederlands "Corpus of Spoken Dutch" (CGN; Dutch Language Union, 2004). The format by Laufer and Nation (1999) was used; that is, for each item, a meaningful sentence was presented with the target word omitted, except for its first letter. When alternative words beginning with the same letter could also be appropriately used, more letters were given.

The frequency of occurrence was purposefully varied to sample words that theoretically differ in difficulty, with low-frequency words hypothesized to be more difficult than high-frequency words. Other characteristics that varied across these items, but that theoretically should not affect item difficulty, were length of the carrier sentences, the lowest frequency of occurrence of the other words within the carrier sentences, and the number of letters for the target word that were already given in the prompt. Table 1 shows the descriptive statistics of the 90 items with respect to the four item characteristics.

Results and summability. Correct and incorrect answers were scored, applying extreme leniency toward spelling mistakes. If the spelling of a response was incorrect but the most likely pronunciation was the same as the correct answer, the item was scored as correct. All inflectional variants of a word were also scored as correct to avoid measuring morphosyntactic knowledge in the vocabulary test.

Summability was measured using the online tool (https://sites.google.com/view/summability). This web application was built using Shiny (Chang et al., 2016) in RStudio Team (2015). In this application, .txt and .sav files can be uploaded with items as columns and participants as rows (or vice versa). We checked the box "answers are dichotomous" (hence, forcing the use of tetrachoric correlations to calculate the covariance matrix). Summability turned out to be .57. (Cronbach's alpha of this 90-item vocabulary test, with the same participants, was .97.) Checking the

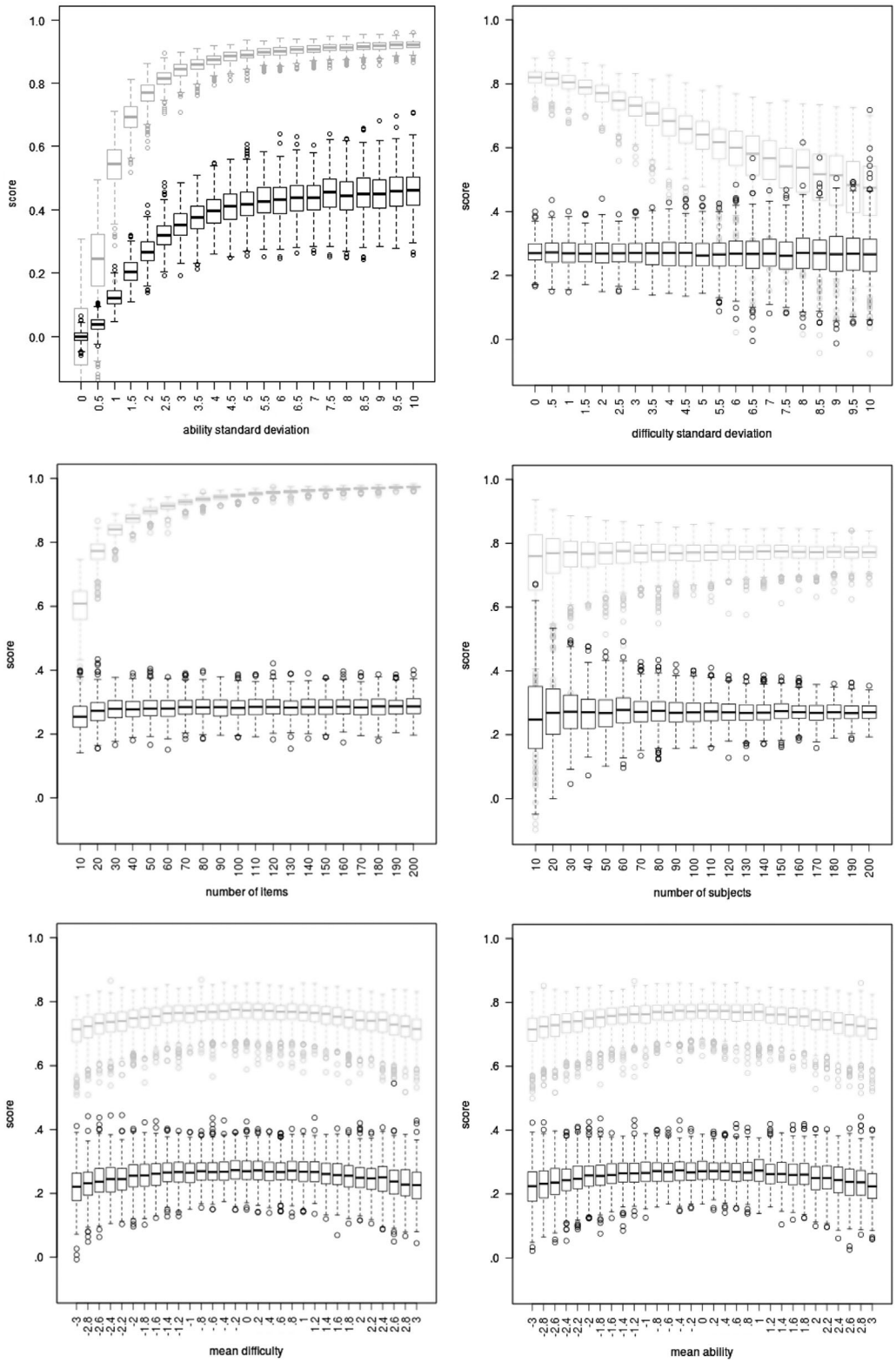

FIGURE 2. Summability (black) and Cronbach's alpha (gray) as a function of various parameters in a simulated mixed Rasch model, in which half of the items follow one Rasch model and the other half follow an independent Rasch model. All boxplots are based on 500 computer simulations.

"Summability-if-item-deleted" box, it turned out that deleting some items would lead to slightly higher summability indices ($s = .59$). For instance, item 17 in the test read "Ze vond het een goed i___ om met de auto te gaan." (*She thought it was a good i___ to take the car)*. The English word "idea" was often given as answer and was marked incorrect, because it cannot be seen as a spelling mistake of the Dutch correct word "idee." This means that transfer from English was marked incorrect, but that potential transfer from French ("idée") was marked as correct. Therefore, in addition to

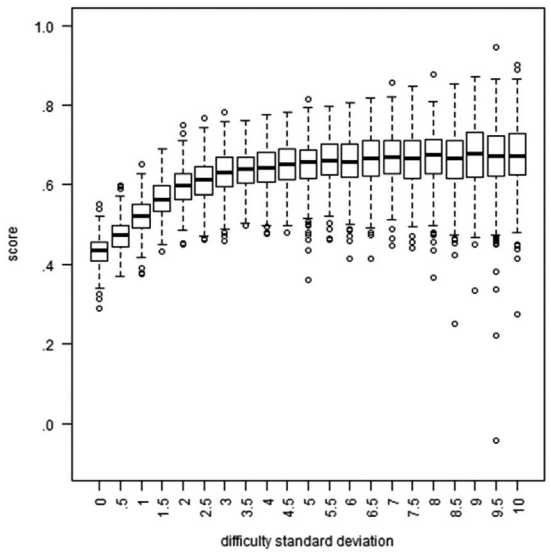

FIGURE 3. Loevinger's H as a function of the variation in item difficulty in a simulated Rasch model. All boxplots are based on 500 computer simulations.

measuring Dutch word knowledge, this item would have given participants with knowledge of French an advantage above participants with knowledge of English. Therefore, we could consider removing this item from the test. Removing it will make the test higher in summability plus we have shown an external rationale why the item may not test productive vocabulary knowledge in a valid manner.

In addition to calculating the summability of a test, we ascertained to what extent the theorized varying item difficulty is related to post hoc calculated item difficulties. Therefore, item difficulty was calculated according to classical test theory, with the total number of correct answers for each item divided by the total number of answers (N = 266). The fifth row of Table 1 shows the descriptive statistics of found item difficulty for these 90 items. Table 2 shows the Pearson correlations between item difficulty and the four item characteristics.

In addition to performing single correlations, we also investigated the predictive value of the item characteristics together in a multiple regression analysis. It was found that frequency of occurrence was the strongest predictor of calculated item difficulty, and none of the other predictors could significantly improve the model (additional R^2 maximally 2.3%, $p > .16$). Hence, only the item characteristic that was identified beforehand as theoretically relevant to predict item difficulty (frequency of occurrence of the words the candidates needed to produce) was a significant predictor of actual item difficulty. None of the other item characteristics could explain any additional variance.

Table 1. Descriptive Statistics of Item Characteristics (and Item Difficulty) of Productive Vocabulary Test (90 Items)

	Mean	*SD*	Range
Number of letters already given	1.71	.66	1–3
Number of words in carrier sentence	15.04	4.74	7–28
(log[a]) word frequency	4.27	1.43	2.83–10.30
(log[a]) frequency of lowest frequency word	4.30	1.94	0–9.4
Item difficulty	.58	.20	.23–.92

[a]Natural logarithm.

Table 2. Pearson Correlations Between Found Item Difficulty and Item Characteristics

	Number of Letters	Number of Words in Carrier Sentence	(log[a]) Word Frequency	(log[a]) Frequency of Lowest Frequency Word
Item difficulty	−.092	−.260	.660	.063
Number of letters		.101	−.220	−.005
Number of words in carrier sentence			−.274	−.349
(log)[a] word frequency				.213

[a]Natural logarithm.

To summarize, we ascertained first that the vocabulary test was an internally consistent test according to high summability, and secondly we showed that there is evidence for the test to be valid, as the separate mini-tests (items) in the test varied in difficulty in a way that was theorized beforehand. Additionally, an item that led to lower summability of the test as a whole was discarded after verifying that this item was indeed an invalid mini-test for the assumed construct (it partly measured language background rather than vocabulary knowledge of Dutch).

Discussion and Conclusion

We have proposed summability as a general measure for the proportion of the total variability measured by a test (or questionnaire) that is captured by the sum score. It measures homogeneity/internal consistency of the items, and unidimensionality in the direction of the sum score. If a test has low summability, much of what is measured by the individual items is lost in the summarization. Tests with low summability could have very noisy items, could measure multiple unrelated constructs, or could have many mismatched items. We pose therefore that a moderate to high summability is a prerequisite for a good test that claims to measure one attribute.

We describe summability as a measure related to internal consistency following Cronbach's use of that term. Summability is a generalization of the average correlation coefficient used by Cronbach (1951) as a measure for internal consistency. This measure is reported by standard software (e.g., SPSS), but often overlooked by practitioners. Based on a novel interpretation, we have generalized this measure to tests with items that are not standardized and to binary outcomes. Summability is also an improved version and a new interpretation of Loevinger's H, which was proposed before as a measure of homogeneity. Summability generalizes Loevinger's H to continuous items. For dichotomous items, it is more stable across tests with more or less variation in item difficulty.

As we have shown in the example of the vocabulary test, reporting a high summability is a prerequisite for a valid test, but it is not sufficient. Summability measures to what

extent all items in the test can be seen as (equal) mini-tests of the same construct, but additional analyses need to be carried out to ascertain to what extent the outcome of the test indeed measures the intended construct. Establishing a strong relation between a priori theorized item difficulty and post hoc found item difficulty, as we did for the productive vocabulary test, is one of the other means to show test validity (e.g., Chapelle, Enright & Jamieson, 2010; Embretson & Gorin, 2001).

Summability is based on a simple decomposition of the covariance matrix of the items and has an interpretation in terms of explained variation similar to an adjusted R^2 in regression. As a measure, it is very general and can be used with items on different scales. It is a descriptive measure that is complementary to model-based analyses, such as factor analysis or Rasch models. If a test has low summability, researchers should look at item-rest correlations and calculate summability-if-item-deleted. Additionally, a factor analysis may be employed to find noisy or mismatched items, or to find out whether a multidimensional construct has been measured. If summability is high, this suggests that a Rasch model would explain the data very well. The advantage of summability over these models, however, is that it gives a single quality measure for the test's sum score, independent of any model.

Summability should replace Cronbach's alpha in its popular and incorrect use as a measure of internal consistency. Internal consistency is consistency between the items of the test, which is important if the items are to be summarized by a sum score. Summability does not have the well known drawback of Cronbach's alpha that it increases dramatically with the test length. Summability is stable in sample size, test length, average difficulty of items, and average ability of subjects, as well as variation in difficulty of items, as we have shown in simulations. Summability should also be used in other situations in which Cronbach's alpha is currently used as a measure of internal consistency, such as in ratings by different judges and survey research with a number of questions for each construct in the survey.

We have given some indication of summabilities that can be considered high or low, but more experience has to be gained on summabilities of tests in various fields before definite recommendations can be given. Definitely, the same criteria that are used for Cronbach's alpha cannot be used for summability, and a summability of 50% should already be considered high in most fields. We invite researchers to calculate summabilities for their tests and to report the results to us, so that a corpus of experience can be obtained. We hope that the online tool (https://sites.google.com/view/summability), which is easy to use, will indeed lead to a fast accumulation of knowledge of what constitutes "high," "moderate," and "low" summability.

Finally, we remark that, although summability is related to unidimensionality and that a summability near 1 is only achievable for a test measuring a unidimensional construct, summability is essentially just a measure of the amount of variation retained when reducing a test to its sum score. This can actually be of great interest for tests explicitly designed to measure multidimensional constructs, such as an overall language proficiency test, which would measure abilities of reading, writing, listening, and speaking. Even within these four skills, subskills can (and usually are) measured. Low summability of the test as a whole would suggest that much information on student performance is discarded by summarizing the test into a single score.

References

Andrich, D. (1988). *Rasch models for measurement*. Sage University Paper Series on Quantitative Applications in the Social Sciences (Vol. 07–068). Newbury Park, CA: Sage.

Bland, J. M., & Altman, D. G. (1997). Statistics notes: Cronbach's alpha. *BMJ*, *314*, 572. https://doi.org/10.1136/bmj.314.7080.572

Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer-Verlag.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, *3*, 296–322.

Chang, W., Cheng, J., Allaire, J. J., Xie, Y., & McPherson, J. (2016). *Shiny: Web Application Framework for R*. R package version 0.13.0. http://CRAN.R-project.org/package=shiny, retrieved on June 1, 2016.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice*, *29*(1), 3–13.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*(1), 98–104.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.

Davenport, E. C., Davison, M. L., Liou, P., & Love, Q. U. (2015). Reliability, dimensionality, and internal consistency as defined by Cronbach: Distinct albeit related concepts. *Educational Measurement: Issues and Practice*, *34*(4), 4–9.

De Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R., & Hulstijn, J. H. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition*, *34*, 5–34.

Dutch Language Union. (2004). *Corpus of spoken Dutch*. Retrieved May 1, 2005, from http://lands.let.ru.nl/cgn.

Embretson, S., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement*, *38*, 343–368.

Green, S. B., Lissitz, R. W., & Mulaik, S. A. (1977). Limitations of coefficient alpha as an index of test unidimensionality. *Educational and Psychological Measurement*, *37*, 827–838.

Guttman, L. L. (1950). The basis for scalogram analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction: Studies in social psychology in World War II* (pp. 60–90). New York, NY: Wiley.

Haertel, E. H. (2006). Reliability. In Brennan, R. L. (Ed.), *Educational measurement* (4th ed., pp. 65–111). Westport, CT: American Council on Education/Praeger.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and ltems. *Applied Psychological Measurement*, *9*(2), 139–164.

Hulstijn, J. H., Schoonen, R., de Jong, N. H., Steinel, M. P., & Florijn, A. (2012). Linguistic competences of learners of Dutch as a second language at the B1 and B2 levels of speaking proficiency of the Common European Framework of Reference for Languages (CEFR). *Language Testing*, *29*(2), 203–221.

Larson-Hall, J. (2010). *A guide to doing statistics in second language research using SPSS*. New York, NY: Routledge.

Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing*, *16*(1), 33–51.

Loevinger, J. (1948). The technic of homogeneous tests compared with some aspects of "scale analysis" and factor analysis. *Psychological Bulletin*, *45*, 507–529.

Mokken, R.J. (1971). *A theory and procedure of scale analysis*. New York, NY: DeGruyter.

Nunnally, J. (1978). *Psychometric methods*. New York, NY: McGraw-Hill.

Pallant, J. (2007). *SPSS survival manual: A step-by-step guide to data analysis using SPSS version 15*. New York, NY: McGraw Hill.

RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA. URL http://www.rstudio.com/, retrieved June 1, 2016.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420–428.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*, 107–120.
Sijtsma, K. (2015). Delimiting coefficient α from internal consistency and unidimensionality. *Educational Measurement: Issues and Practice*, *34*(4), 10–13.
Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, *3*, 271–295.
Warrens, M. J. (2008). On association coefficients for 2×2 tables and properties that do not depend on the marginal distributions. *Psychometrika*, *73*, 777–789.
Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cornbach's α, Revelle's β, and McDonalds ω_H: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, *70*(10), 123–133.