Psychon Bull Rev (2016) 23:750–763
DOI 10.3758/s13423-015-0968-3

THEORETICAL REVIEW

Use of internal consistency coefficients for estimating reliability of experimental task scores

Samuel B. Green[1] · **Yanyun Yang**[2] · **Mary Alt**[3] · **Shara Brinkley**[1] · **Shelley Gray**[1] · **Tiffany Hogan**[4] · **Nelson Cowan**[5]

Published online: 6 November 2015

Abstract Reliabilities of scores for experimental tasks are likely to differ from one study to another to the extent that the task stimuli change, the number of trials varies, the type of individuals taking the task changes, the administration conditions are altered, or the focal task variable differs. Given that reliabilities vary as a function of the design of these tasks and the characteristics of the individuals taking them, making inferences about the reliability of scores in an ongoing study based on reliability estimates from prior studies is precarious. Thus, it would be advantageous to estimate reliability based on data from the ongoing study. We argue that internal consistency estimates of reliability are underutilized for experimental task data and in many applications could provide this information using a single administration of a task. We discuss different methods for computing internal consistency estimates with a generalized coefficient alpha and the conditions under which these estimates are accurate. We illustrate use of these coefficients using data for three different tasks.

Keywords Reliability · Coefficient alpha · Split-half reliability · Generalized coefficient alpha

✉ Samuel B. Green
samgreen@asu.edu

[1] Arizona State University, Social Sciences Building, 951 S Cady Mall, P.O. Box 873701, Tempe, AZ 85287-3701, USA

[2] Florida State University, Tallahassee, FL, USA

[3] University of Arizona, Tucson, AZ, USA

[4] MGH Institute of Health Professions, Charlestown, MA, USA

[5] University of Missouri, Columbia, MO, USA

When conducting research with paper-and-pencil measures, researchers frequently report reliability coefficients and most often coefficient alpha, an internal consistency estimate of reliability (Cortina, 1993; Hogan, Benjamin, & Brezinski, 2000; Peterson, 1994). From our experience, reliability estimates are much less frequently reported for experimental task measures, those on which individuals respond to relatively novel stimuli presented across multiple trials. In support of our hypothesis, we reviewed articles published in Volume 21 and the Issues 1 through 3 of Volume 22 in *Psychonomic Bulletin and Review* (PB&R) and Volumes 71 through 83 in the *Journal of Memory and Language* (JML). Of the 245 relevant articles in PB&R, 14 of them reported reliability coefficients: nine internal consistency reliability coefficients (alpha and split-half) and six interrater reliabilities. Of the 94 relevant articles in JML, 11 reported reliability coefficients: five internal consistency reliability coefficients (alpha, split-half, and split-third), one test–retest reliability, and five interrater reliabilities. These results are consistent with our hypothesis that reliabilities are infrequently presented for experimental task scores.

One possible explanation for these results is that it is unnecessary to assess the reliability of task scores in experimental studies. However, this explanation is incorrect. Low reliability negatively impacts effect size, power of hypothesis tests, and replicability of results across studies, regardless of the design or method of analysis (e.g., Cleary, Linn, & Walster, 1970; Humphreys & Drasgow, 1989; LeBel & Paunonen, 2011). Thus, knowledge of reliability of task scores can help in understanding results within a study as well as differences between studies.

Another explanation is that researchers need not compute reliabilities for tasks within their studies, but rather can rely on reliability coefficients reported in previous studies. However, it is difficult to infer, based on reported coefficients in one set

of studies, how reliable data are in other studies, particularly for bespoke experimental tasks. Reliability may vary with changes in the number of trials administered on a task, the properties of the stimuli, the amount of time allotted to respond to stimuli, and the order of task administration. Reliability also may vary depending on the age and the ability of respondents as well as the assessment environment (Thompson, 2003). Thus, it is very useful to have insight into the reliability of a task measure as it is administered within one's own study.

The focus of our paper is on a third explanation for why reliability coefficients are infrequently reported for experimental task measures: a perceived lack of methods for computing coefficients that are relatively easy to apply and are accurate in their assessment of reliability. In terms of ease of use, internal consistency coefficients require only a single administration of a measure and thus can be computed based on the data collected in one's study. Nevertheless, it is not always obvious whether internal consistency coefficients are appropriate and how to compute them to yield an accurate assessment of reliability. The purpose of our paper is to address the use of internal consistency reliability coefficients for experimental task measures.

In the following sections, we define reliability and discuss a variety of internal consistency coefficients that could be applied with experimental task measures and their underlying assumptions. We focus on a particular family of internal consistency coefficients, those using a general formulation of coefficient alpha. We illustrate issues that arise in research practice by computing these coefficients on computer-generated data as well as data collected on second graders who were administered tasks designed to assess working memory and word learning. Throughout our presentation, we argue that internal consistency coefficients can be viable estimates of reliability for experimental task measures if they are applied in a judicious fashion. In our conclusion section, we briefly discuss internal consistency reliability coefficients other than those emphasized in this paper.

Definition of reliability

We begin by defining reliability of experimental task scores within the framework of classical test theory (CTT). For an individual *i*, a task observed score, ξ_i, is the sum of a task true score, τ_i, and a task error score, ε_i; that is,

$$\xi_i = \tau_i + \varepsilon_i \,. \tag{1}$$

It is assumed that τ_i and ε_i are uncorrelated, and the mean of the error scores in the population of participants is zero. Then, the task observed variance is a sum of the task true variance and task error variance; that is, $\sigma_\xi^2 = \sigma_\tau^2 + \sigma_\varepsilon^2$. Reliability is defined as a ratio of the task true score variance to the task observed score variance, $\sigma_\tau^2/\sigma_\xi^2$. As demonstrated in CTT, reliability also can be shown to be equal to the correlation between scores on a task and a parallel task $\left(\rho_{\xi\xi'}\right)$; that is,

$$\rho_{\xi\xi'} = \sigma_{\xi\xi'}/\sigma_{\xi'}\sigma_\xi = \sigma_\tau^2/\sigma_\xi^2, \tag{2}$$

where $\sigma_{\xi\xi'}$ is the covariance between the task and the parallel task scores, σ_ξ is the standard deviation of the task scores, and $\sigma_{\xi'}$ is the standard deviation of the parallel task scores.

Internal consistency coefficients and coefficient alpha

In this section, we present various internal consistency coefficients. Essentially, all the coefficients are special cases of coefficient alpha. The distinction among them is based on how trials on a task are combined together into parts (or splits) of a task before computing coefficient alpha. In describing applications of coefficient alpha with experimental task measures, we substitute the terms *trials* and *tasks* for *items* and *tests*, which are the terms traditionally used in describing reliability with paper-and-pencil tests.

Coefficient alpha as a family of internal consistency coefficients

For any task, we might contemplate computing coefficient alpha on trial data. However, potentially a coefficient alpha can be calculated on scores for any set of components or splits of a task (Raju, 1977). Thus, there is a family of coefficient alphas that can be computed for any task. Some coefficient alphas in this family are likely to be better estimates of reliability than others. In this section, we describe the family of coefficient alphas. In subsequent sections, we discuss how to choose among the alphas in the family to obtain the best estimate of reliability.

We begin by presenting the general formulation for coefficient alpha. Prior to computing any one coefficient in the family of coefficient alphas, trials are combined together to create *K* splits or components of a measure. For simplicity, we consider only equal-sized splits among the *N* trials. (For unequal-size splits, see Raju, 1977, and Warrens, 2014.) In computing coefficient alpha, scores are computed for each split as well as for the task. A split score is any statistic that summarizes the performance across trials for a split, whereas a task score is the sum of the split scores. Next, we compute the

mean of the covariances $\left(\overline{\sigma}_{\text{split,split}'}\right)$ between the $K(K-1)/2$ pairs of splits and the variance of the task (σ^2_{Task}). We now can substitute into the general formula for coefficient alpha, denoted $\alpha_{\text{split }1/K}$:

$$\alpha_{\text{split }1/K} = \frac{K^2\ \overline{\sigma}_{\text{split,split}'}}{\sigma^2_{\text{Task}}}. \tag{3}$$

$\alpha_{\text{split }1/K}$ is an estimate of the reliability of task scores (i.e., summed scores across splits) or a linear transformation of these scores (e.g., mean of split scores). In computing $_{\text{split }1/K}$, we have a choice of how to partition the N trials into K splits. By far, the most popular method is to split tasks at the trial level (i.e., $K = N$). $\alpha_{\text{split }1/K}$ now can be reformulated as coefficient alpha for trial data:

$$\alpha_{\text{trial}} = N^2\frac{\overline{\sigma}_{\text{trial,trial}'}}{\sigma^2_{\text{Task}}}. \tag{4}$$

$\overline{\sigma}_{\text{trial,trial}'}$ is the mean covariance between all pairs of trial scores, and σ^2_{Task} is the variance of the summed trial scores. With α_{trial}, we are splitting a task into the most number of components.

Alternatively, we can split a task into two halves (i.e., $K = 2$), the fewest number of components. In this case, we divide trials on a task into halves, compute scores on the two halves, and then calculate a reformulated $\alpha_{\text{split }1/K}$:

$$_{\text{split half}} = \frac{4\ \sigma_{\text{half,half}'}}{\sigma^2_{\text{Task}}}. \tag{5}$$

$\sigma_{\text{half,half}'}$ is the covariance between halves, and σ^2_{Task} is the variance of the sum of the scores across the two halves.[1]

It also is possible to compute coefficient alpha for splits when K is between 2 and N. For example, we could divide the trials on a task into thirds, compute scores on each third, and calculate a coefficient alpha for split-third data:

$$_{\text{split third}} = \frac{9\,\overline{\sigma}_{third,third'}}{\sigma^2_{Task}}. \tag{6}$$

$\overline{\sigma}_{third,third'}$ is the mean covariance among three pairs of thirds, and σ^2_{Task} is the variance of the summed scores across the three thirds.

[1] A split-half coefficient also may be calculated by computing a correlation between two halves of a measure and then applying the Spearman-Brown prophesy formula to estimate the reliability of the whole measure. This approach requires the two halves to be parallel, which is more restrictive than the essential tau equivalence assumption for the split half coefficient using coefficient alpha. Split-half coefficients using the two approaches yield similar results if the difference between variances for the two halves is small or moderate (Warrens, 2015).

Assumptions underlying coefficient alpha

The assumptions underlying coefficient alpha are associated with the observed split scores on a task in the population. For example, the assumptions involve trial scores for α_{trial} or scores on the halves for $\alpha_{\text{split half}}$. The assumptions for the split scores are as follows: (a) An observed score for a split is a sum of true and error scores. (b) The true scores are within an additive constant of each other across splits (i.e., the essential tau equivalence assumption). Within a factor analytic model, a single factor underlies all splits, the loadings on this factor are the same for all splits, and the split scores differ by an additive constant across individuals. (c) The error scores for each split have a mean of zero, are uncorrelated with the true scores, and are uncorrelated with error scores of other splits (i.e., the uncorrelated errors assumption).

In the literature, researchers have focused primarily on the essential tau equivalence assumption and, to a lesser extent, on the uncorrelated errors assumption (e.g., Cortina, 1993; Green, 2003; Green & Hershberger, 2000; Miller, 1995; Osburn, 2000; Raykov, 1997). Coefficient alpha is an underestimate of reliability to the extent that splits are not essentially tau equivalent and an overestimate to the degree that the error scores are positively correlated across trials. We want to split task measures so that the effects of violating these assumptions are minimized. Thus, the crucial decision is the choice among the many ways to split a task.

How to split a task to minimize violation of assumptions

The most popular method for computing reliability is coefficient alpha for item or trial data, α_{trial} (Cronbach, 1951; Cronbach & Shavelson, 2004). A major reason for its popularity is that it is easy to determine. It involves a single administration of a measure and requires minimal judgment about splitting a measure; that is, each split consists of a single trial. Of course, ease of use does not imply that it yields the most accurate estimate of reliability. As an alternative to α_{trial}, we could compute a split-half coefficient (Spearman, 1910) or some other split-1/K coefficient (e.g., $\alpha_{\text{split-third}}$). Researchers are faced with a difficult judgment in computing $\alpha_{\text{split }1/K}$: the choice of a particular split of a measure prior to the computation of coefficient alpha.

How to split a task to minimize violation of the essential tau equivalence assumption

The choice between α_{trial} and $\alpha_{\text{split half}}$ can be framed in terms of their underlying assumptions. α_{trial} is an appropriate choice if all trials are essentially tau equivalent (and errors are

uncorrelated). In the population, if trials are essentially tau equivalent, all $\alpha_{\text{split half}}$ are the same and equal to a measure's reliability; however, split-half estimates will vary across splits in a sample due to sampling error. Thus, it would be best to average these split-half coefficients to yield the most stable estimate of reliability. Cronbach (1951) proved that coefficient alpha for trials is the mean of all possible split-half coefficients for a measure. Thus, coefficient alpha for trial data is preferred if the assumptions underlying alpha are met at the trial level.

In contrast, the split-half coefficient is likely to be the preferred estimate of reliability if trials are not essentially tau equivalent. In this case, $\alpha_{\text{split half}}$ for a measure will vary from split to split in the population, although none may exceed the reliability, assuming errors are uncorrelated. Some split-half coefficients may yield accurate estimates of reliability if the halves are essentially tau equivalent, nearly accurate estimates if the halves are approximately essentially tau equivalent, or poor estimates if the halves are far from essentially tau equivalent. Thus, the decision about how to split trials into halves is important because some split-half coefficients are likely to be better estimates of reliability than others.

Some researchers (Callender & Osburn, 1977, 1979; Osburn, 2000) have recommended an empirical strategy for choosing the split that obtains the largest split-half coefficient. They argued that the maximal $\alpha_{\text{split half}}$ least underestimates reliability when the essential tau equivalence assumption is violated. There are two problems with this recommendation (Thompson, Green, & Yang, 2010). First, in choosing the largest $\alpha_{\text{split half}}$, we are capitalizing on the chance characteristics of a sample and thus are likely to report an overestimate of the true reliability, unless the sample is very large. Second, we are ignoring that coefficients also can be affected by violations of the uncorrelated errors assumption. Thus, we prefer to make the split decision based on our understanding of a task, which is ideally grounded in past research, and then to offer, if possible, empirical support for this decision based on the current sample data.

The initial choice between α_{trial} and $\alpha_{\text{split half}}$ as well as the subsequent choice among different split-half coefficients may not be important if violation of essential tau equivalence has minimal effect on the accuracy of these coefficients as estimates of reliability. Based on a number of studies (e.g., Feldt & Qualls, 1996; Green & Yang, 2009a; Raykov, 1997; Zinbarg, Revelle, Yovel, & Li, 2005), α_{trial} can yield poor estimate of reliability if multiple trials on the general factor have weak loadings and other trial(s) have stronger loadings, the task has few trials, and the task is multidimensional. Overall, it appears that α_{trial} may be a reasonably accurate estimate of reliability if the task contains a moderate to large number of trials (perhaps 20 or more) and the trials have been selected to be included on the task as a function of their statistical properties.

For simplicity, we have focused on the way to minimize violation of the essential tau equivalence assumption for α_{trial} and $\alpha_{\text{split half}}$. However the approach is the same for any $\alpha_{\text{split 1/K}}$. We combine trials together to create split scores that are as equivalent as possible.

How to split a task to minimize violation of the uncorrelated errors assumption

Psychometricians warned about difficulties with meeting the uncorrelated errors assumption for internal consistency reliability estimates (Cronbach & Shavelson, 2004; Guttman, 1945; Rozeboom, 1966). We suspect there may be fewer ways to violate this assumption with task measures in comparison with traditional tests (Green & Hershberger, 2000; Green & Thompson, 2003), but this is speculation. Errors may be correlated if different subsets of trials on a task are associated with different stimuli. For example, one set of trials may be associated with one particular visual stimulus and another set with a different visual stimulus (e.g., Steinberg & Thissen, 1996; Wainer & Kiely, 1987; Yen, 1993). Alternatively, respondents might tend to do better (or worse) on a trial if they performed well (poorly) on the previous trial, regardless of their level of skills on the task. In statistical terms, an autoregressive or moving averages process may underlie the performance across trials (Green & Hershberger, 2000). Positive correlations between errors across trials yield an inflated coefficient alpha and an overestimate of reliability. The effect can be quite strong (e.g., Fleishman & Benson, 1987; Komaroff, 1997; Maxwell, 1968; Miller, 1995; Raykov, 1998), but the actual size of the effect is unclear in practice.

We want to minimize the effects of violations of the uncorrelated errors assumption in splitting a task into components. For example, if four trials are associated with each of five stimuli on a 20-trial task, a researcher might combine across the four trials for each stimulus and compute a split-fifth coefficient alpha. With autoregressive or moving averages effects, trials that are closer to each other in the administration of a task are likely to have higher correlations than those farther apart, even if all trials are measuring the same skills. To minimize this effect, we create splits such that trials associated with different splits are as far apart as possible in the administration of the task (Green & Yang, 2005).

Examples of computing internal consistency coefficients for tasks

In the following sections, we use examples to illustrate issues that arise in research practice in computing internal consistency reliability coefficients on task scores. These examples are based on data collected on second graders with typical

development from Arizona, Nebraska, and Massachusetts. The students completed the Comprehensive Assessment Battery for Children (Cabbage et al., 2015; Gray et al., 2015), which included tasks assessing working memory and word learning. The tasks were completed over 4 days. To maintain the interest of the students, the tasks were developed as computer games involving pirates and monsters. We discuss computation of internal consistency reliability for three tasks: learning referent, location span running, and classic Stroop tasks. For the classic Stroop task, we augment the experimental results by computer generating and analyzing simulated Stroop data.

Choosing a coefficient for a learning referent task

In this section, we illustrate how to define the scores of interest for a task (i.e., the focal task scores) and how to split the task so that the resulting coefficient alpha is a reasonable reliability estimate of the focal task scores. The measure in our example is a learning referent task for assessing a child's ability to learn the names of four novel sea monsters. For any one child, two monsters are randomly assigned two-syllable names, and the other two monsters are randomly assigned four-syllable names. For any trial, children see all four monsters on a computer screen, hear a name, and touch the monster on the screen to indicate which monster goes with the presented name. They receive immediate feedback by being given a virtual coin if correct or a virtual rock if incorrect. The first block of trials was considered fast-mapping in that students completed two trials for each of the four names. In each of the three subsequent blocks, children complete 60 trials, with 15 trials for each name. After completing all blocks, children use their earned coins to shop at a virtual ice-cream shop. For our study, 160 second-grade students with typical development completed the task.

Given that children may or may not respond similarly to learning two-syllable versus four-syllable words, we chose to compute separate scores for the two types of names. Accuracies (i.e., proportion of correct trials) were computed for each name type within a block. These accuracy scores represent how well they have learned the task for a block but do not represent the ability of children to learn names. Thus, the reliabilities of these scores were not of primary interest, although they may have diagnostic value.

We considered three focal variables to assess speed in learning for a child: (1) change in accuracies from Block 2 to Block 4; (2) the mean difference in accuracies between adjacent blocks across Blocks 2, 3, and 4; and (3) the least squares slope predicting accuracies from block number across Blocks 2, 3, and 4. For all three variables, accuracies were computed across the two names within each name type. As we show next, these three variables are essentially comparable.

The least squares regression slope is

$$\beta_i = \frac{\sum_{b=2}^{4}\left(X_b-\overline{X}\right)\left(Y_{ib}-\overline{Y}_i\right)}{\sum_{b=2}^{4}\left(X_b-\overline{X}\right)^2}, \tag{7}$$

where Y_{ib} is an accuracy scores for respondent i in block b, and X_b is a block number. Equation 7 simplifies to

$$\begin{aligned}\beta_i &= \frac{(2-3)\left(Y_{i2}-\overline{Y}_i\right)+(3-3)\left(Y_{i3}-\overline{Y}_i\right)+(4-3)\left(Y_{i4}-\overline{Y}_i\right)}{(2-3)^2+(3-3)^2+(4-3)^2}\\ &= \frac{Y_{i4}-Y_{i2}}{2} = \frac{(Y_{i3}-Y_{i2})+(Y_{i4}-Y_{i3})}{2}.\end{aligned} \tag{8}$$

Based on Eq. 8, the least squares slope (i.e., focal variable 3) is mathematically equivalent to the mean difference in accuracies between adjacent blocks (i.e., focal variable 2), and perfectly correlated with the change in accuracy from block 2 to block 4 (i.e., focal variable 1).

We chose to conceptualize the focal task scores as slopes across Blocks 2, 3, and 4, although these slopes can be redefined as difference scores between blocks, as just demonstrated. To be consistent, we will define split scores for computing reliability coefficients as slopes across Blocks 2, 3, and 4. A requirement in splitting a measure is that the focal variable is a sum of the split scores or a linear combination of the summed scores. In addition, the scores for the splits should be essentially tau equivalent and their errors independent of each other. We considered a number of ways to split the task but narrowed our choice to one of the three possibilities. These three methods for computing reliabilities of the slopes across blocks are described below and presented pictorially in Fig. 1.

Split by trial number This method assessed consistency across trials to estimate reliability. It was computed as follows: (1) Accuracies were computed as a mean across the 2 names, separately for the 2 name types and 15 trials within the 3 blocks. Thus, 90 accuracies were computed for each child (90 = 2 name types × 15 trials × 3 blocks). (2) A slope was calculated for accuracies across the 3 blocks, separately by the 2 name types and 15 trials; therefore, 30 slopes were calculated for each child (30 = 2 name types × 15 trials). (3) Finally, $\alpha_{\text{split-15th}}$ was computed for each name type. These reliabilities were based on the consistency among the 15 slopes defined by trials for a name type.

Split by name This method assessed consistency across names to estimate reliability. It was computed as follows: (1)

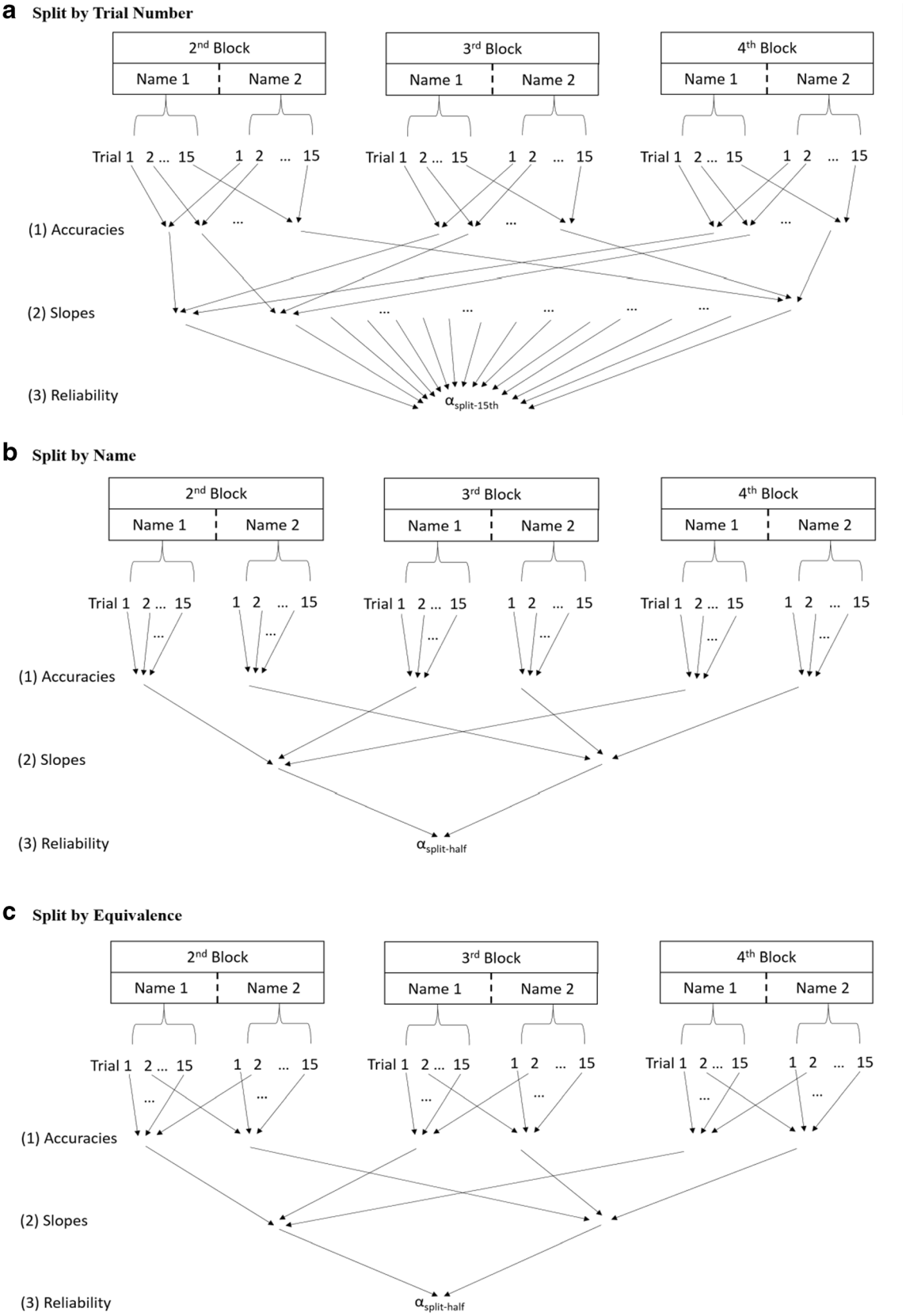

Fig. 1 Pictorial representation of three splitting methods for a learning referent

Accuracy was computed as a mean across the 15 trials, separately for the 2 words within each of the 2 word types for the 3 blocks. Thus, 12 accuracies were computed for each child (12 = 2 names × 2 name types × 3 blocks). (2) A slope was calculated for accuracies across the 3 blocks, separately by the 2 names within the 2 name types; therefore, 4 slopes were calculated for each child (4 = 2 names × 2 name types). (3) Finally, $\alpha_{\text{split-half}}$ was computed for each name type. These reliabilities were based the consistency between the 2 slopes defined by the 2 names for a name type.

Split for equivalence This method assessed consistency across halves created to be as similar as possible in terms of names and trial sequence. It was computed as follows: (1) Accuracy was computed as a mean across the 15 trials, balancing the effects of trial sequence and word; accuracy was calculated separately for the 2 equated halves within each of the 2 word types for the 3 blocks, producing 12 accuracies for a child (12 = 2 equated halves × 3 blocks × 2 name types). The accuracy for one equated half was computed as a mean across the 8 odd-numbered trials for the first name and the 7 even-numbered trials for the second name within a name type. The accuracy for the other equated half was computed as a mean across the 8 odd-numbered trials for the second name and the 7 even-numbered trials for the first name for a name type. (2) A slope was calculated for accuracies across the 3 blocks, separately for the 2 equated halves within the 2 name types; therefore, 4 slopes were calculated for a child (4 = 2 equated halves × 2 name types). (3) Finally, $\alpha_{\text{split-half}}$ was computed for each name type. These reliabilities were based the consistency between the 2 slopes defined by equated halves for a name type.

We weighed the positives and the negatives associated with the three splitting methods in assessing the reliabilities of the slopes. As required for computing a coefficient alpha, the sums of the slopes for the splits for all three splitting methods are linearly related to the focal task scores, the slopes based on the accuracies across the three blocks. It is straightforward to prove this conclusion in that the slopes can be redefined in terms of difference scores. In other applications, it may be difficult to demonstrate mathematically that the sum of the split scores is linearly related to the focal task scores. An alternative to a derivation is to compute the correlation between the sum of the split scores and the focal task scores. If the resulting correlation is 1.0, the required condition has been met for the analyzed data, although not necessarily in general. It should be noted that the correlation would have been less than 1.0 for the third method if one half had included the odd numbered trials and the other half had included the even numbered trials because the number of odd and even numbered trials differed.

We next consider the violation of the essential tau equivalence assumption. In constructing the task, names were selected within the two-syllable pair and within the four-syllable pair so that they would be as equivalent as possible. Nevertheless, any two names have different structures and associations and are likely to be learned with differential difficulty across children. This lack of equivalence of the names within the two-syllable and the four-syllable pairs is likely to produce underestimates of reliability using the second splitting method. Another characteristic about the task that could affect equivalence of the splits for the second method is trial number. Although the accuracies are likely to differ across trials, it is less apparent whether the relative speeds in learning are likely to differ across the 15 trials. In the last method we created splits that were relatively balanced within split with respect to name and trial number. Accordingly, we believed a priori that this method would yield the most equivalent split.

We last consider the violation of the uncorrelated error assumption. We would argue that the first and third splitting methods are most vulnerable with regard to this assumption in that the splits are a function of trial number, and thus there is the potential for a sequential effect. However, we do not think the concern is a major one because the occurrence of adjacent numbered trials does not imply adjacent trials in the task. In addition, any sequential effect should be minimized, given the scores computed for splits are slopes across blocks. Overall, we chose the three split methods because they were judged to produce relatively appropriate splits. However, of the three methods, we preferred the third approach because splits based on it were more likely to meet the essential tau equivalence assumption.

As shown in Table 1, the reliabilities varied by splitting methods. As expected, the reliability estimates were highest,

Table 1 Internal Consistency Reliabilities for the Learning Referent Task

Variable	Splitting Method		
	Split by trial number	Split by name	Split for equivalence
Two-syllable names			
Block 2	.88	.81	.91
Block 3	.92	.91	.93
Block 4	.93	.93	.95
Slopes	.73	.72	.80
Four-syllable names			
Block 2	.90	.86	.93
Block 3	.92	.87	.93
Block 4	.92	.84	.90
Slopes	.81	.70	.84

Note. Coefficient alphas were computed not only for the slopes but also for the accuracies of the blocks. To calculate these alphas for blocks, we followed the same first step as described in computing alphas for slopes. Then, we calculated (1) alphas based on the consistency between accuracies of the 15 trials created by splitting the task by trial number, (2) alphas based on the consistency between accuracies of the halves created by splitting the task by name within name type, and (3) alphas based on the consistency between accuracies of the halves created to maximize equivalence

with one exception, for the split method for equivalence. We would argue that because this method came closest to meeting the essential tau equivalence assumption, it yielded the most accurate reliability estimates. On the other hand, the reliability estimates based on halves that differed by names tended to produce the lowest reliability. Although the names were chosen to minimize differences in responses, any two names are likely to yield different results, and therefore our findings were not surprising.

The focal task scores of slopes were reliable, although not highly reliable. This result was not unexpected in that the slope scores are essentially difference scores (see Eq. 8), and difference scores are more unreliable in many applications than the scores that go into them (Rogosa & Willett, 1983). Consistent with this literature and as shown in Table 1, the block scores evidenced greater reliability than the slopes.

Choosing a coefficient for a location span running task

We next present a brief example using a location span running task. The reason for presenting this task is to illustrate that splits that are not routinely used in practice, such as split thirds or split quarters, may make better sense for some tasks. For the location span running task, the game was to help direct a pirate to buried treasure by remembering where a series of arrows pointed in sequence from the last to the first location. At the beginning of each trial, a black dot appeared at the center of the screen, followed by sequentially presented arrows (5, 6, 7, or 8 arrows). The arrows radiated out from the black dot and pointed at discrete locations at one of eight equidistant angles in a clock-like pattern. The number of arrows that appeared in a trial was randomly determined so that children could not anticipate span length. After a sequence of arrows was presented, eight red dots appeared in a circular pattern around the screen to show the possible locations where arrows could have pointed. The children were asked to touch the red dots indicating the locations where arrows had pointed, sequentially from the last to the first location. The task included three trials for each of the four span lengths. The dependent variable was the mean number of locations correctly identified across all trials. One-hundred and fifty-four children completed this task for the analyses.

To maximize the chances of meeting the essential tau equivalence assumption, we split the task into thirds, with scores for any third being the mean number of locations correctly identified across four trials of different span lengths. This approach is a master plan for splitting the task rather than yielding a unique split. We considered three unique splits within this master plan, as presented in Table 2. For the first splitting method, the 1st, 2nd, and 3rd thirds consisted of the first trials for the four span lengths, the second trials for the four span lengths, and the third trials for the four span lengths, respectively. We were interested in this splitting method because the thirds based on this method should be the least equivalent among methods within the master plan. More specifically, the thirds within this splitting method differ dramatically in terms of when they were presented in the task: earliest for the 1st third, next earliest for the 2nd third, and latest for the 3rd third. In contrast, the second splitting method in the table should yield thirds with greater equivalence, and the third splitting method even greater equivalence.

As hypothesized and as shown in Table 2, the coefficient alphas differed as a function of the degree of equivalence of thirds across the splitting methods. We would argue that .94 represents the best estimate of reliability for the focal variable, not simply because it is the highest value among the three coefficients but because the splits for this coefficient best meet the assumptions underlying alpha.

Choosing a coefficient for the stroop test

In our final example, we discuss the choice of internal consistency reliability coefficients for the Stroop color-word task (or, more simply, the Stroop test), which was designed to assess inhibition. We present this example because the scoring of the Stroop test can produce data that could be inconsistent with the essential tau equivalence assumption. Results are presented for computer-simulated Stroop data as well as data from our experimental study.

With the Stroop test, one of four words describing colors (e.g., "red," "blue," "yellow," or "green") appears on a computer screen. These words are displayed in one of four colors (e.g., red, blue, yellow, or green). Individuals are instructed to name the displayed color of the word on the screen as quickly as possible. The task includes both congruent and incongruent trials, with the order of presentation randomly or semirandomly determined. The presented word and its displayed color are the

Table 2 Trial Numbers for the Four Span Lengths Involved in the Creation of Split Thirds for Three Splitting Methods and Their Resulting Split-Third Coefficients

Splitting method	Spans of 1st third				Spans of 2nd third				Spans of 3rd third				$\alpha_{\text{split-third}}$
	5	6	7	8	5	6	7	8	5	6	7	8	
Method 1	1	1	1	1	2	2	2	2	3	3	3	3	.86
Method 2	1	1	3	3	2	2	2	2	3	3	1	1	.92
Method 3	2	2	1	3	3	1	2	2	1	3	3	1	.94

same for congruent trials and different for incongruent trials. Multiple dependent variables can be computed for the Stroop test: proportion of trials answered correctly (accuracies), the mean response times for the trials answered correctly (RTs), and an index to assess inhibition (e.g., difference in mean RTs between conditions).

In splitting the Stroop test to compute a coefficient alpha, the mean RTs for the various splits can be based on different numbers of trials, depending on how many trials are answered correctly by respondents. In addition, the number of trials answered correctly in the various splits is likely to vary across respondents. This process is not built into the CTT model underlying coefficient alpha. Because the essential tau equivalence assumption is violated, coefficient alpha may underestimate the true reliability. We analyzed computer generated data to assess the bias of coefficient alpha in estimating the true reliability for Stroop-like data. We also considered whether violation of the uncorrelated errors assumption has an effect on the bias due to variability in the number of trial across splits.

Analysis of simulated stroop RT data In Appendix A, we describe the methods used to generate RT data and to calculate reliabilities on these data. For simplicity, we generated data for a 16-trial task representing a single condition (e.g., incongruent trials). In the simulation, we varied the percentage of trials answered correctly: 100 % or 80 %. The choice of trials to be answered incorrectly was randomly determined. We also varied whether the uncorrelated errors assumption was violated or not. If the assumption was violated, the strongest correlation was for errors between adjacent trials. We computed coefficient alphas based on seven different splits of the task: α_{trial} for trial-level data, $\alpha_{split\ half}$ for four different splits halves of the task, and $\alpha_{split\ quarter}$ for three different split quarters of the task. The splits for $\alpha_{split\ half}$ and $\alpha_{split\ quarter}$ are shown in Table 3.

Population-level analysis Initially, we evaluated coefficient alphas at the population level. We approximated these population coefficients by generating data for 5,000,000 simulees.

Because the data were computer generated, we could compute the true reliability. The population results are presented in Table 4. In the first column of this table are the results when all trials were answered correctly and the uncorrelated errors assumption was met. The true reliability based on the simulated data (.716) was identical to the true reliability based on the mathematical model underlying the data, as presented in Appendix A. This result acts as a validity check for the simulation computer program. As expected given, the method used to generate the data, the population alphas, regardless of how the task was split, were accurate estimates of the true reliability.

When only 80 % of the trials were answered correctly with no violation of the uncorrelated errors assumption (in the third column of Table 4), the true reliability was lower (.665) because the observed task scores (mean RT for trials answered correctly) were based on fewer number of trials. The split-quarter coefficient alphas marginally underestimated the true reliability, whereas the split-half coefficient alphas were slightly better. At first glance, the results for α_{trial} for these conditions appear surprising in that violation of the essential tau equivalence assumption should yield lower-bound estimates. However, the inaccuracy of α_{trial} was due to "missing data" in the calculation of alpha. All trials had to be answered correctly for the data for a simulee to be included in the calculation of α_{trial}. The result was that the alpha was based on only 3 % of the 5,000,000 simulees. It is not surprising that the α_{trial} for this subset of simulees was identical to the alpha for the condition with 100 % accuracy because both were computed for data that were generated comparably.

When the uncorrelated error assumption was violated and all trials were answered correctly, the coefficient alphas, regardless of how the task was split, were overestimates of the true reliability. The overestimation was minimal for some splits and dramatic for other splits. The overestimation was greater to the extent that adjacent trials were in different splits. When only 80 % of the trials were answered correctly, the effect of violation of the uncorrelated errors assumption was minimized for split-half and split-quarter alphas. Essentially, the correlation between errors for different trials decreased

Table 3 Trials for Different Splits for $\alpha_{split\ half}$ and $\alpha_{split\ quarter}$

Split name	1st split	2th split	3rd split	4th split
Split halves				
By 1 s	1,3,5,7,9,11,13,15	2,4,6,8,10,12,14,16		
By 2 s	1,2,5,6,9,10,13,14	3,4,7,8,11,12,15,16		
By 4 s	1,2,3,4,9,10,11,12	5,6,7,8,13,14,15,16		
By 8 s	1,2,3,4,5,6,7,8	9,10,11,12,13,14,15,16		
Split quarters				
By 1 s	1,5,9,13	2,6,10,14	3,7,11,15	4,8,12,16
By 2 s	1,2,9,10	3,4,11,12	5,6,13,14	7,8,15,16
By 4 s	1,2,3,4	5,6,7,8	9,10,11,12	13,14,15,16

Table 4 Population Values of α_{trial}, $\alpha_{split\ half}$, and $\alpha_{split\ quarter}$ for Different Splits (Percent of Simulees Used in Computation of Coefficients in Parentheses, if less than 100 %)

Coefficient	100 % Accuracy		80 % Accuracy	
	Uncorrelated errors	Correlated errors	Uncorrelated errors	Correlated errors
True reliability				
$\rho_{\xi\xi'}$	.716	.633	.665	.623
Split into items for α_{item}				
α_{trial}	.716	.757	.716 (3 %)	.738 (1 %)
Split halves for $\alpha_{split\ half}$				
$\alpha_{split\ half\ by\ 1s}$	.716	.828	.660 (≈ 100 %)	.616 (≈ 100 %)
$\alpha_{split\ half\ by\ 2s}$	.716	.760	.660 (≈ 100 %)	.615 (≈ 100 %)
$\alpha_{split\ half\ by\ 4s}$	.716	.692	.660 (≈ 100 %)	.615 (≈ 100 %)
$\alpha_{split\ half\ by\ 8s}$	.716	.653	.660 (≈ 100 %)	.616 (≈ 100 %)
Split quarters for $\alpha_{split\ quarter}$				
$\alpha_{split\ quarter\ by\ 1s}$	.716	.787	.648 (99 %)	.601(98 %)
$\alpha_{split\ quarter\ by\ 2s}$	.716	.721	.649 (99 %)	.601(98 %)
$\alpha_{split\ quarter\ by\ 4s}$	.716	.672	.648 (99 %)	.600 (98 %)

Note. If a simulee responded incorrectly for a trial, the simulated RT was essentially treated as missing data. Data for a simulee were included in the computation of alpha only if scores were available for all splits. The implication was most dramatic with α_{trial} for the 80 % accuracy condition; only 1 % to 3 % of the simulees were included in the analysis to compute these coefficients (as shown in parentheses in the table)

when 20 % of the trial scores were excluded from the computation of the coefficient alphas. As before with 80 % accuracy, the α_{trial} was problematic because it was based on data for only a very small percentage of the examinees (1 %).

Sample-level analysis We also generated 1,000 samples with 100 simulees to evaluate coefficient alpha values for sample data. The means of the coefficient alphas across the 1,000 samples are presented in Table 5. Overall, the results for the mean sample values were similar to those at the population level. Regardless of condition, the sample coefficient alphas were slightly negatively biased estimates of their population counterparts. Although not shown in the table, the instability of the sample estimates (evaluated by the standard deviation

Table 5 Mean α_{trial}, $\alpha_{split\ half}$, and $\alpha_{split\ quarter}$ for Different Splits

Coefficient	100 % Accuracy		80 % Accuracy	
	Uncorrelated errors	Correlated errors	Uncorrelated errors	Correlated errors
True reliability				
$\rho_{\xi\xi'}$	.716	.633	.665	.623
Split into items for α_{item}				
α_{trial}	.709	.751	Insufficient Data	Insufficient Data
Split halves for $\alpha_{split\ half}$				
$\alpha_{split\ half\ by\ 1s}$	.709	.824	.651	.605
$\alpha_{split\ half\ by\ 2s}$	.707	.752	.649	.602
$\alpha_{split\ half\ by\ 4s}$	.710	.685	.655	.606
$\alpha_{split\ half\ by\ 8s}$	.710	.645	.652	.607
Split quarters for $\alpha_{split\ quarter}$				
$\alpha_{split\ quarter\ by\ 1s}$	.708	.802	.639	.588
$\alpha_{split\ quarter\ by\ 2s}$	.710	.714	.640	.589
$\alpha_{split\ quarter\ by\ 4s}$	.710	.665	.642	.592

Note. If a simulee responded incorrectly for a trial, the simulated RT was essentially treated as missing data. Data for a simulee were included in the computation of alpha if scores were available for all splits. For the two 80 % accuracy conditions, the number of simulees with complete item data was judged insufficient for computing α_{trial} (eight or fewer for the 1,000 samples). In comparison, the number of simulees with complete data was much greater for computing $\alpha_{split\ half}$ (99 or 100 for the 1,000 samples) and $\alpha_{split\ quarter}$ (96 to 100 for the 1,000 samples)

of the sample coefficients) was greater if splits contained more trials.

It is important to note that we do not present results for α_{trial} for the 80 % accuracy conditions because these alphas were based on datasets with very few simulees (eight or fewer simulees in any one sample). Although not presented, the alphas based on these extremely small samples were wildly inaccurate. These results indicate why trial-level coefficient alphas are not computed for Stroop data in practice.

Overall the simulation data indicated that if the Stroop test is split appropriately, coefficient alpha can yield accurate estimates of reliability. We should add that we generated simulated data using alternative parameter values for the underlying model and had similar conclusions based on their results. However, further work is necessary to evaluate the accuracy of coefficient alphas for alternative underlying models.

Experimental data For the experimental data, children were presented 12 congruent trials and 12 incongruent trials, with the order of presentation randomly determined for each child. We computed mean RTs for trials answered correctly for congruent and incongruent conditions as well as the differences between these mean RTs. Mean RTs were considered valid only if 6 of the 12 trials were answered correctly for each condition. The sample size was reduced from 161 to 156 as a function of this restriction.

A variety of internal consistency coefficients for Stroop RT data are presented in Table 6. A child's data were used to compute a coefficient alpha using a splitting method for a condition only if at least half of the trials were answered correctly for all splits: at least 3 correct trials for each half in computing an $\alpha_{split\ half}$, at least two trials for each third in computing an $\alpha_{split\ third}$, and all correct trials to compute α_{trial}.

As we see in the first row of Table 6, the reliability coefficients based on trial-level data are problematic. As previously discussed with the simulated data, a child's data were not included in the computation of α_{trial} unless that child answered all trials correctly. For example, for the incongruent condition, α_{trial} was computed based on the RTs for only 81 of the 156 children. α_{trial} based on a reduced sample size is likely to be a biased estimate of the true reliability and cannot be trusted. The other coefficient alphas reported in Table 6 were based on sample sizes that were close to the total analysis sample size of 156 and thus should have been minimally susceptible to the sample bias problem of α_{trial}. However, a similar problem could arise with other alphas if research participants would make more errors than those in our sample.

It is interesting to note that different splitting methods had only a minor effect on the values of coefficient alpha for the incongruent and congruent conditions. The alphas were mostly in the low .70s for the incongruent condition, whereas the alphas were in the high .50s and low .60s for the congruent condition. Alphas also were computed based on the differences in RTs between conditions (last column of Table 6). These difference scores, rather than the RTs for the incongruent and congruent conditions, are frequently used to infer inhibition effects. The alphas for these difference scores were variable but generally indicated poor reliability. Ostensibly, we would conclude that the difference scores demonstrated unsatisfactory reliability in our sample, and a greater number of trials should be considered in future research with similar sampled children. This conclusion is bolstered by previous research that concluded that differences between incongruent and congruent RTs on the Stroop test are unreliable based on test–retest coefficients, although the RTs for the incongruent

Table 6 Internal Consistency Reliabilities of RTs (*N*s in parentheses) for the Classic Stroop Task

Coefficient	Split[a]	Incongruent	Congruent	Difference
α_{trial}	1/2/3/4/5/6/7/8/9/10/11/12	.76 (81)	.61 (132)	.21 (68)
$\alpha_{split\ half}$	1,3,5,7,9,11 / 2,4,6,8,10,12	.73 (155)	.61 (156)	.40 (155)
$\alpha_{split\ half}$	1,2,5,6,9,10 / 3,4,7,8,11,12	.73 (156)	.59 (156)	.38 (156)
$\alpha_{split\ half}$	1,2,3,7,8,9 / 4,5,6,10,11,12	.75 (156)	.56 (156)	.34 (156)
$\alpha_{split\ half}$	1,2,3,4,5,6 / 7,8,9,10,11,12	.67 (155)	.54 (156)	.14 (155)
$\alpha_{split\ third}$	1,4,7,10 / 2,5,8,11 / 3,6,9,12	.71 (153)	.63 (156)	.36 (153)
$\alpha_{split\ third}$	1,2,7,8 / 3,4,9,10 / 5,6,11,12	.74 (154)	.57 (156)	.33 (154)
$\alpha_{split\ third}$	1,2,3,4 / 5,6,7,8 / 9,10,11,12	.70 (155)	.57 (156)	.23 (155)

Note. Reliabilities were based on a sample of 156 children. For the congruent condition, 146 children responded correctly to 100 % of the trials, and 10 children responded correctly to at least 83 % but less than 100 % of the trials. For the incongruent conditions, 83 children responded correctly to 100 % of the trials; 64 children responded correctly to at least 83 % but less than 100 % of the trials; and 9 children responded correctly to at least 58 % but less than 83 % of the trials

[a] Commas are used to separate trials within a split, and slashes are used to separate splits

and congruent RTs are reliable (e.g., Jensen, 1965; Siegrist, 1997; Strauss, Allen, Jorgensen, & Cramer, 2005).

Conclusion

Unfortunately it is risky to rely on past studies that reported reliability estimates given that reliability is a function of the characteristics of the task, the conditions under which it is administered, and the type of respondents. Thus, it would be advantageous to estimate reliability based on the same data used to reach substantive conclusions. Internal consistency estimates of reliability for task measures are ideal for this purpose because they can provide this information using a single administration of a task. Of course, they are only ideal if they yield relatively accurate estimates of reliability.

We have discussed different methods for computing internal consistency estimates with a generalized coefficient alpha and the conditions under which these estimates are accurate. We have argued that experimenters should carefully consider different methods to split their task and choose the split that should yield the most accurate estimate of reliability. They also may consider computing multiple coefficient alphas based on different splits of a task to assess whether the calculated values of these coefficients are consistent with their understanding of the measure. We demonstrated through the examples the thought process underlying the choice of coefficient alpha and the confidence gained in choosing a reliability estimate if the results are consistent with one's predictions. If the coefficient alphas yield low values, as with the Stroop difference scores, the experimenter should be skeptical about the reliability of the task measure.

As with any data collection and analysis method, we need to be judicious in the way that we collect and analyze data in the computation of coefficient alpha. However, with methodological care, coefficient alpha should provide helpful information in understanding the reliability of experimental task scores. Most importantly, as currently practiced, we have set a low bar for coefficient alpha: Is it better than guessing the reliability based on previous studies? On the basis of our analyses, we would suggest that this bar has been met.

We presented coefficient alpha as a general method for computing reliability, regardless of the choice of splits. The advantage of this method is that it allows for a unified approach for computing coefficients, regardless of the number of splits and with a minimal amount of statistical estimation complexities. However, it is important to recognize that coefficients other than alpha are available. For example, the Angoff-Feldt coefficient for two components and the Feldt-Gilmer coefficient for three or more components (Feldt & Charter, 2003) do not require essentially tau equivalent components but rather the less restrictive assumption of congeneric components (i.e., allowing for differences among loadings on a single underlying factor). Alternatively, reliability coefficients can be computed based on results of factor analytic models, which can be unidimensional or multidimensional (McDonald, 1999; Zinbarg, Yovel, Revelle, & McDonald, 2006). In addition, the coefficients based on factor analytic models can take into account component scores that are ordinal (Green & Yang, 2009b).

We would like to see further discussion of reliability coefficients for task scores. In particular, there is little data to evaluate the sensitivity of reliability coefficients to changes in task characteristics, respondents, and administration conditions. Vacha-Haase, Henson, and Caruso (2002) discuss a methodology for evaluating sensitivity of reliability coefficients, which they refer to as reliability generalization. Also, a typology of task measures could be suggested that would link the kind of internal consistency reliability coefficients to the type of task. Potentially, we would learn most about the applicability of internal consistency reliability coefficients for task measures by applying them in experimental and field research.

Author note This work was supported by funding from the National Institutes of Health NIDCD Grant #R01 DC010784. We are deeply grateful to the staff, research associates, school administrators, teachers, children, and families who participated. Key personnel included (in alphabetical order) Shara Brinkley, Katy Cabbage, Gary Carstensen, Cecilia Figueroa, Karen Guilmette, Trudy Kuo, Bjorg LeSueur, Annelise Pesch, and Jean Zimmer. Many students also contributed to this work including (in alphabetical order) Genesis Arizmendi, Lauren Baron, Alexander Brown, Nora Schlesinger, Nisha Talanki, and Hui-Chun Yang.

Appendix A. Describing simulation and analysis of RTs for stroop task

In this appendix, we describe the methods used to generate simulated RT data and to calculate coefficient alphas and true reliability for these data.

Generation method

Within CTT, the observed score is equal to the true score plus an error score for an individual i on trial t: $\xi_{it} = \tau_i + \varepsilon_{it}$. We generated the true and error scores and then summed them to obtain observed scores, the simulated RTs. To generate a true score on a trial for an individual, we generated scores that were exponentially distributed with a parameter, λ; the mean and variance of this distribution were $1/\lambda$ and $1/\lambda^2$, respectively. We simulated variability in true scores across individuals by generating normally distributed scores, μ_i, with a mean of μ and a variance of $\sigma^2_{\mu_i}$. Note that μ_i is constant across trials for an individual so that the scores on trials are tau equivalent. Errors were generated to be normally distributed with a mean of 0 and a variance of $\sigma^2_{\varepsilon} = \sigma^2 + 1/\lambda^2$. To create observed

scores, we summed the true and error scores. These observed scores were ex-Gaussian distributed across items for an individual. The observed scores can alternatively be defined as a sum of two independent random variables, one of which is distributed exponentially and the other normally. The exponentially distributed variable has a single parameter, λ, and the normally distributed variable has a mean of μ_i and a variance of σ^2. The mean and variance of trial simulated RTs for an individual are $\mu_{\xi_{it}} = \mu_i + 1/\lambda$ and $\sigma^2_{\xi_{it}} = \sigma^2 + 1/\lambda^2$, respectively. The distribution of observed scores has a skew of $(2/\sigma_X^{3/2}\lambda^3)(1+2/\sigma_X^2\lambda^2)^{-3/2}$.

Based on the generation model, the reliability of simulated RTs for a trial is

$$\rho_{\xi\xi'} = \frac{\sigma^2_\tau}{\sigma^2_\tau + \sigma^2_\varepsilon} = \frac{\sigma^2_{\mu_i}}{\sigma^2_{\mu_i} + \left(\sigma^2 + 1/\lambda^2\right)}, \tag{A.1}$$

and the reliability for experimental task scores if all trials are answered correctly is

$$\rho_{\xi\xi'} = \frac{N\sigma^2_{\mu_i}}{N\sigma^2_{\mu_i} + \left(\sigma^2 + 1/\lambda^2\right)}. \tag{A.2}$$

It is also straightforward to prove that the reliability for an experimental task is equal to coefficient alpha for our generation model:

$$\rho_{\xi\xi'} = \frac{N\sigma^2_\tau}{N\sigma^2_\tau + \left(\sigma^2 + 1/\lambda^2\right)} = N^2\frac{\bar{\sigma}_{trial,trial'}}{\sigma^2_{Task}}. \tag{A.3}$$

Because the trial data are tau equivalent, coefficient alpha is equal to the population reliability for all other equal splits of trials.

It is more difficult to assess the accuracy of coefficient alpha if all trials are not answered correctly or if the errors are correlated. For our simulation, we used a random uniform number generator to define the probability of answering the trial correctly. We also simulated correlated errors by creating an autoregressive error structure. More specifically, for an individual, the error at trial *t* is a function of the error at time *t* – 1 with an autoregressive coefficient of β plus an independent error at time *t* times a coefficient to maintain stationarity; that is,

$$\varepsilon_{it} = \beta\,\varepsilon_{i(t-1)} + \left(1-\beta^2\right)^{1/2}\varepsilon'_{it}. \tag{A.4}$$

Data for our simulation

For our simulations, we generated data for a 16-trial task using the following parameters: μ = 600, $\sigma^2_{\mu i}$ = 6400, λ = .005, and σ^2 = 625. These values are similar to those considered by Ratcliff (1993), although he did not consider reliability of simulated RTs and thus did not decompose his observed scores into true and error scores. Based on these parameters and all trials being answered correctly, the population reliability for a trial is

$$\rho_{\xi\xi'} = \frac{N\sigma^2_{\mu_i}}{N\sigma^2_{\mu_i} + \left(\sigma^2 + 1/\lambda^2\right)} = \frac{6400}{6400 + \left[625 + 1/(.005)^2\right]} = .136,$$

and the population reliability for the task is

$$\rho_{\xi\xi'} = \frac{N\sigma^2_T}{N\sigma^2_T + \left(\sigma^2 + 1/\lambda^2\right)} = \frac{N\sigma^2_T}{N\sigma^2_T + \left(\sigma^2 + 1/\lambda^2\right)} = 0.716.$$

In the simulation, we varied the percent of trials answered correctly: 100 % and 80 %. We also varied the autoregressive coefficient with β being set at 0 or .2 to assess the effects of violating the uncorrelated errors assumption.

We examined reliability both at the population level and sample level. We generated data for 5,000,000 simulees to approximate reliability values at the population level, whereas we generated 1,000 samples with 100 simulees to evaluate reliability values at the sample level.

Assessing reliabilities

Because we generated the data, we were able to compute a very accurate approximation of the true reliability. At the population level, we generated retest scores for the 18 trials of the task by maintaining the same true scores for the test and retest but creating different error scores. The true reliability then was determined by computing a correlation between the test and the retest for the 5,000,000 simulees.

We assessed the accuracy of coefficient alphas based on seven different splits of the task: an α_{trial} for trial-level data, $\alpha_{split\ half}$ for four different half-splits of the task, and $\alpha_{split\ quarter}$ for three different quarter-splits of the task. The splits for $\alpha_{split\ half}$ and $\alpha_{split\ quarter}$ are shown in Table 3 in the body of the paper. These coefficient alphas were computed both at the population and sample levels and compared to the true reliability to assess their accuracy.

References

Cabbage, K. L., Brinkley, S., Gray, S., Alt, M., Cowan, N., Green, S., … Hogan, T. P. (2015). *The comprehensive assessment battery for children: Working memory* (CABC-WM). Manuscript submitted for publication.

Callender, J. C., & Osburn, H. G. (1977). A method for maximizing split-half reliability coefficients. *Educational and Psychological Measurement, 37,* 819–825.

Callender, J. C., & Osburn, H. G. (1979). An empirical comparison of coefficient alpha, Guttman's lambda-2, and msplit maximized split-half reliability estimates. *Journal of Educational Measurement, 16,* 89–99.

Cleary, T. A., Linn, R. L., & Walster, G. W. (1970). Effect of reliability and validity on power of statistical tests. *Sociological Methodology*, 130–138.

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

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334.

Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement, 64*(3), 391–418.

Feldt, L. S., & Charter, R. A. (2003). Estimating the reliability of a test split into two parts of equal or unequal length. *Psychological Methods, 8,* 102–109.

Feldt, L. S., & Qualls, A. L. (1996). Bias in coefficient alpha arising from heterogeneity of test content. *Applied Measurement in Education, 9,* 277–286.

Fleishman, J., & Benson, J. (1987). Using LISREL to evaluate measurement models and scale reliability. *Educational and Psychological Measurement, 47*(4), 925–939.

Gray, S., Green, S., Alt, M., Hogan, T., Kuo, T., Brinkley, S., & Cowan, N. (2015). *The structure of working memory in young children and its relation to nonverbal intelligence.* Manuscript submitted for publication.

Green, S. B. (2003). A coefficient alpha for test-retest data. *Psychological Methods, 8,* 88–101.

Green, S. B., & Hershberger, S. L. (2000). Correlated errors in true score models and their effect on coefficient alpha. *Structural Equation Modeling, 7,* 251–270.

Green, S. B., & Thompson, M. S. (2003). Structural equation modeling in clinical research. In M. C. Roberts & S. S. Illardi (Eds.), *Methods of research in clinical psychology: A handbook* (pp. 138–175). London: Blackwell.

Green, S. B., & Yang, Y. (2005). *K-Split coefficient alpha*. Presented at Annual Meeting of the American Educational Research Association, Montreal, Canada.

Green, S. B., & Yang, Y. (2009a). Commentary on coefficient alpha: A cautionary tale. *Psychometrika, 74*(1), 121–135.

Green, S. B., & Yang, Y. (2009b). Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha. *Psychometrika, 74*(1), 155–167.

Guttman, L. A. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10,* 255–282.

Hogan, T. P., Benjamin, A., & Brezinski, K. L. (2000). Reliability methods: A note on the frequency of use of various types. *Educational and Psychological Measurement, 60,* 523–531.

Humphreys, L. G., & Drasgow, F. (1989). Some comments on the relation between reliability and statistical power. *Applied Psychological Measurement, 13*(4), 419–425.

Jensen, A. R. (1965). Scoring the Stroop test. *Acta Psychologica, 24,* 398–408.

Komaroff, E. (1997). Effect of simultaneous violations of essential τ-equivalence and uncorrelated error on coefficient α. *Applied Psychological Measurement, 21,* 337–348.

LeBel, E. P., & Paunonen, S. V. (2011). Sexy but often unreliable: The impact of unreliability on the replicability of experimental findings with implicit measures. *Personality and Social Psychology Bulletin, 37,* 570–583.

Maxwell, A. E. (1968). The effect of correlated errors on estimates of reliability coefficients. *Educational and Psychological Measurement, 28,* 803–811.

McDonald, R. P. (1999). *Test theory: A unified approach*. Mahwah: Erlbaum.

Miller, M. B. (1995). Coefficient alpha: A basic introduction from the perspectives of classical test theory and structural equation modeling. *Structural Equation Modeling, 2,* 255–273.

Osburn, H. G. (2000). Coefficient alpha and related internal consistency reliability coefficients. *Psychological Methods, 5,* 343–355.

Peterson, R. A. (1994). A meta-analysis of Cronbach's coefficient alpha. *Journal of Consumer Research, 21*(2), 381–391.

Raju, N. S. (1977). A generalization of coefficient alpha. *Psychometrika, 42,* 549–565.

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin, 114,* 510–532.

Raykov, T. (1997). Scale reliability, Cronbach's coefficient alpha, and violations of essential tau-equivalence with fixed congeneric components. *Multivariate Behavioral Research, 32,* 329–353.

Raykov, T. (1998). Coefficient alpha and composite reliability with interrelated nonhomogeneous items. *Applied Psychological Measurement, 22,* 375–385.

Rogosa, D. R., & Willett, J. B. (1983). Demonstrating the reliability the difference score in the measurement of change. *Journal of Educational Measurement, 20*(4), 335–343.

Rozeboom, W. W. (1966). *Foundations of the theory of prediction*. Homewood: Dorsey.

Siegrist, M. (1997). Test-retest reliability of different versions of the Stroop test. *The Journal of Psychology, 131*(3), 299–306.

Spearman, C. (1910). Correlation calculated with faulty data. *British Journal of Psychology, 3,* 271–295.

Steinberg, L., & Thissen, D. (1996). Uses of item response theory and the testlet concept in the measurement of psychopathology. *Psychological Methods, 1,* 81–97.

Strauss, G. P., Allen, D. N., Jorgensen, M. L., & Cramer, S. L. (2005). Test-retest reliability of standard and emotional Stroop tasks an investigation of color-word and picture-word versions. *Assessment, 12*(3), 330–337.

Thompson, B. (Ed.). (2003). *Score reliability: Contemporary thinking on reliability Issues*. Newbury Park: Sage.

Thompson, B. L., Green, S. B., & Yang, Y. (2010). Assessment of the maximal split-half coefficient to estimate reliability. *Educational and Psychological Measurement, 70,* 232–251.

Vacha-Haase, T., Henson, R. K., & Caruso, J. (2002). Reliability generalization: Moving toward improved understanding and use of score reliability. *Educational and Psychological Measurement, 62,* 562–569.

Wainer, H., & Kiely, G. (1987). Item clusters and computerized adaptive testing: The case for testlets. *Journal of Educational Measurement, 24,* 189–205.

Warrens, M. J. (2014). On Cronbach's alpha as the mean of all possible *k*-split alphas. *Advances in Statistics*. ID 742863.

Warrens, M. J. (2015). A comparison of reliability coefficients for psychometric tests that consist of two parts. *Advances in Data Analysis and Classification*, 1–14.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30,* 187–214.

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and Mcdonald's ω_h: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika, 70*(1), 123–133.

Zinbarg, R., Yovel, I., Revelle, W., & McDonald, R. (2006). Estimating generalizability to a universe of indicators that all have one attribute in common: A comparison of estimators for omega. *Applied Psychological Measurement, 30,* 121–144.