Estimating Reliability with Panel Data: Alternatives to Cronbach's Alpha and Why We Need Them

Paul Biemer[1], **Sharon Christ[2]**, Christopher Wiesen[3]
[1]RTI, International, RTP, North Carolina
[2]Odum Institute for Research in Social Science, UNC Chapel Hill, Chapel Hill, NC
[3]Odum Institute for Research in Social Science, UNC Chapel Hill

1. Introduction

Scale score measurements (SSM's) are very common in psychological and social science research. As an example, the Child Behavior Checklist (CBCL) is a common SSM for measuring behavior problems in children (see Achenbach, 1991a, 1991b for the version of the CBCL used in this paper). It consists of 118 items on behavior problems, each scored on a 3-point scale: 1 = not true, 2 = sometimes true and 3 = often true of the child. The CBCL Total Behavior Problem Score is an empirical measure of child behavior computed as a sum of the responses to the 118 items. The usefulness of any SSM in data analysis depends in large part on its reliability. An SSM having poor reliability is infected with random errors that obscure the true construct underlying the measure. SSM's having good reliability are relatively free from random error which enhances their validity as an analysis variable (see, for example, Biemer and Trewin, 1997). For example, Biemer and Trewin show that as reliability (R) decreases, the standard errors of estimates of means, totals and proportions increase by the factor $\sqrt{R^{-1}}$. In addition, for simple linear regression, the slope coefficient is biased toward 0 if the explanatory variable is not reliable. Thus, assessing scale score reliability is typically an integral and critical step in the use of SSM's in data analysis.

The most common method for assessing scale score reliability is Cronbach's α (Hogan, Benjamin, & Brezinsky, 2000). A number of software packages for data analysis (for e.g., SAS, SPSS, and STATA) provide subroutines for computing α with relative ease. There are numerous examples in the literature of using α for assessing the reliability of scale scores. One reason for α's ubiquity is that few alternative methods for assessing reliability in cross-sectional studies are available – this despite the fact that α has been criticized in the literature due to the rather strong assumptions underlying its development as an indicator of reliability (see, for example, Bollen, 1989, p.217; Cortina, 1993; Green & Hershberger, 2000; Luke, 2005; Zimmerman & Zumbo, 1993).

It is well-known that α tends to overestimate reliability when the SSM items are subject to inter-item correlated error (Green & Hershberger, 2000; Lucke, 2005; Raykov, 2001; Rae, 2006; Vehkalahti, et al, 2006; Zimmerman, et. al, 1993; Komaroff, 1997). The assumption of uncorrelated inter-item error is violated, for example, if respondents try to respond consistently to the items in scale rather than considering each item independently of the others and providing the most accurate answer to each. For items which are prone to social desirability effects, errors across items may be correlated if respondents force their responses to be more socially acceptable than the truth may seem. Respondents may also respond as they think they should rather than completely honestly, a form of acquiescence bias. These situations tend to induce positively correlated errors which will positively bias α; i.e., reliability as measured by α will appear higher than it truly is.

Cronbach's α can also underestimate reliability if the items in an SSM do not all measure the same construct (Raykov, 1998; Raykov & Shrout, 2002; Komaroff, 1997). For example, an SSM that is intended to measure depression may include some items that instead measure anger or pain. In addition, the questions may be worded in such a way that respondents interpret the questions erroneously and report behaviors or attitudes which are inconsistent with the construct of interest.

For panel data, there are alternatives to α that rely on assumptions that are more easily satisfied in practice. One of these is the simplex estimator of reliability (Wiley and Wiley, 1970). Unlike α, the simplex estimator is a function of the sum score itself rather than individual scale items and, therefore, it accuracy is not affected by inter-item correlated error. When the scale items are subject to correlated error, simplex reliability estimates will tend to be smaller than Cronbach's α which, as noted previously, is inflated. This is not to say that simplex estimates are always more accurate than Cronbach's α since the simplex model assumptions can also be violated. This raises a question for the analyst who computes both estimates: if the estimates differ considerably, which has the greater accuracy (or validity) and should be reported? This question should be address for each application since the model assumptions are satisfied to varying degrees depending on the SSM and the study design.

This paper proposes an approach, referred to as the generalize simplex method, for estimating scale score reliability for panel data under more general assumptions than those required for either α or the simplex estimator. It will be shown that, by imposing parameter restrictions on the model underlying this new estimator, estimates of reliability that are consistent with Cronbach's α, the simplex method or even several other useful simplex-like approaches can be produced. This provides the analyst with a number of options for reporting SSM reliability.

As an example, in situations where its quality can be assured, Cronbach's α may be preferred over more complex estimators of reliability since it is widely used and easy to compute. The generalized simplex method can be used to test whether the assumptions underlying α or several alternative estimators of reliability hold for a particular SSM. In cases where α's assumptions are rejected, our approach provides a process for identifying the simplest method for computing reliability whose quality can be verified by formal tests of significance. In some situations an analyst may prefer to compute the generalize simplex estimate of reliability without testing whether simpler alternatives are available. However, it can be instructive to identify situations where the assumptions underlying α and the traditional simplex model do not hold to inform future uses of these methods.

For example, to the extent that SSM's perform similarly across a range of study settings and designs, testing the assumptions underlying reliability estimation would be quite useful to analysts who contemplate using the same or similar SSM's in other data sets. As an example, if the assumption of uncorrelated errors is rejected for an SSM in one particular study, that should serve as a warning that this assumption may be questionable for this SSM across studies. In some situations, it may be possible to modify the data collection methodology to reduce inter-item correlated error for the SSM. At a minimum, it would forewarn analysts that the use of Cronbach's α for assessing the SSM's reliability is suspect.

The next section briefly reviews the concept of reliability, particularly scale-score reliability, and introduces the notation and models that will be needed for describing the methods. We show that Cronbach's α and the simplex method are essentially special cases of the generalize simplex method which is uses the method of split-halves (Bollen, 1989, p. 213-215). The methodology for testing the assumptions underlying alternative estimates of reliability also developed. In Section 3, we apply this methodology to a number of scale score measures from the National Survey of Child and Adolescent Well-being (NSCAW) to illustrate the concepts and the performance of the estimators.

2. Scale Score Reliability

Observations obtained in a survey or other study are subject to errors which may be attributable to a number of error sources including survey questions, respondents, interviewers and data processing procedures. These error sources impart both systematic and random errors to the measurements. For a particular data item, assume there is a true value, μ_i, for the ith individual in the survey; however, rather than observing μ_i, we observe y_i. The difference $e_i = y_i - \mu_i$ is the measurement error; that is, $y_i = \mu_i + e_i$. For the ith individual, the mean of the e_i's over hypothetical repetitions of the measurement process is the systematic component of error denoted by b_i; i.e., $E(e_i|i) = b_i$. The sum of an individual's true value and this systematic component, i.e. $t_i = \mu_i + b_i$, is called the true score of the individual. It is simply the mean of the hypothetical distribution of responses for an individual. These assumptions lead to the error model

$$y_i = \mu_i + b_i + \varepsilon_i \qquad (1)$$

or equivalently,

$$y_i = t_i + \varepsilon_i \qquad (2)$$

where $\varepsilon_i = y_i - t_i$ and $\mathrm{E}(\varepsilon_i \mid i) = 0$. Define the variance of the ε_i's as

$$\sigma^2_{\varepsilon i} = \mathrm{E}(\varepsilon_i^2 \mid i) \qquad (3).$$

If we further assume that $\operatorname{cov}(\varepsilon_i, \varepsilon_j) = 0$ and $\operatorname{cov}(\varepsilon_i, t_i) = 0$, then the unconditional variance of y_i is given by

$$\begin{aligned}\mathrm{Var}(y_i) &= \mathrm{Var}(t_i) + \mathrm{Var}(\varepsilon_i) \\ &= \sigma_t^2 + \sigma_\varepsilon^2\end{aligned} \qquad (4)$$

Reliability analysis is concerned with the amount of variable error that is present in the process for measuring the true value, μ_i. The reliability ratio is

$$R = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_\varepsilon^2} \qquad (5)$$

(see, for example, Fuller, 1987, p. 3) defined as true score variance divided by the total variance of the measurements (i.e., the sum of true score and random error variance). Reliability is essentially the proportion of total variance that is true score variance. It can also be interpreted as the intraclass correlation coefficient (ICC) among items or the proportion of total variation among items and subjects due to the shared variance (or correlation) of the items. When R is high, we say the measurement process is reliable; i.e., the variation in the measurements is due mostly to the variation in the true scores of individuals in the population. When R is low, we say that the measurement process is unreliable; that is, the variation in the measurements is mostly random error or "noise."

The same concepts can be applied to an SSM (or multi-item scale) which can be defined broadly as any sequence of questions that assesses facets of the same construct to produce a scale score, S. For our purposes, S is defined as the unweighted sum of responses to the questions comprising the SSM. Each item in the scale is assumed to be measured on an ordinal scale (for e.g., a Likert scale) and is an indicator of the same latent construct. If we assume that the measurement errors for the items are uncorrelated (i.e., no inter-item correlated error), the reliability of the score S can be estimated as a function of the inter-item correlations. This is the basis for Cronbach's α method of estimating the reliability of S (Cronbach, 1951).

The next sections describe three models for estimating scale score reliability beginning the with simplest approach, Cronbach's α. A second method, referred to as the simplex method, will then be introduced that can be applied when the same construct is measured at three or more time points or panel waves. Finally, we develop the generalized simplex approach which also requires three or more waves of data. In addition, it assumes that the SSM can be divided into two psychometrically equivalent SSM's using the method of split halves. As we shall see, the α and simplex models are special cases of this generalized simplex model.

2.1 Estimating Reliability Using Cronbach's α

To fix the ideas, a four-item SSM will be assumed initially and subsequently generalized for k >2 items. The assumptions underlying Cronbach's α can be illustrated by the simple factor analysis model in the following model equations:

$$\begin{aligned} y_1 &= \lambda_1 t + \varepsilon_1 \\ y_2 &= \lambda_2 t + \varepsilon_2 \\ y_3 &= \lambda_3 t + \varepsilon_3 \\ y_4 &= \lambda_4 t + \varepsilon_4 \end{aligned} \qquad (6)$$

where $y_j, j = 1,...,4$ denote the responses to the four items for a particular individual, t denotes the true score which is the same for all four indicators and ε_j are random error terms. The subscript, i, denoting the individual has been dropped as a notational convenience. The λ_j's are scaling coefficients to adjust for differences in the scales of measurement among the items.

The model also assumes that the measurement errors, ε_j 's are uncorrelated between items; i.e., $\mathrm{cov}(\varepsilon_j, \varepsilon_{j'}) = 0$ for any two items j and j'. In addition, Cronbach's α assumes that the four measurements are parallel; that is, $\lambda_j = 1$ and $Var(\varepsilon_j) = \sigma_\varepsilon^2$, for all j. This implies that all four items are measured using the same scale of measurement and are subject to the same error distribution.

Now generalizing to k items, define the scale score, S, for a k-item scale as $S = \sum_j y_j$. From (6), it follows that

$$Var(S) = k^2\sigma_t^2 + k\sigma_\varepsilon^2 \qquad (7)$$

The first term on the right side of (7) is the true score variance and the second term is the error variance. Under this model, the reliability of S is given by

$$R = \frac{k^2\sigma_t^2}{k^2\sigma_t^2 + k\sigma_\varepsilon^2} \qquad (8)$$

or

$$R = \frac{\sigma_t^2}{\sigma_t^2 + \frac{\sigma_\varepsilon^2}{k}} \qquad (9)$$

Note from (9) that the error variance component is divided by k, the number of items in the scale which implies that reliability increases as the number of items in the scale increases. Thus, according to the assumptions of Cronbach's α, a 50-item scale will be more reliable than a scale consisting of a subset of k<50 of these items. Failure of this relationship between k and R to hold is evidence that the assumptions underlying Cronbach's α also do not hold.

Under these assumptions, an unbiased estimator of R in (9) is Cronbach's α given by

$$\hat{\alpha} = \left(\frac{k}{k-1}\right)\left(1 - \frac{\sum_{j=1}^{k} \mathrm{var}(y_j)}{\mathrm{var}(S)}\right) \qquad (10)$$

where $\mathrm{var}(y_j)$ is an estimate of $Var(y_j)$ and $\mathrm{var}(S)$ is an estimate of $Var(S)$. For a simple random sample of size n, the an unbiased estimator of $\mathrm{var}(y_j)$ is

$$\mathrm{var}(y_j) = \frac{\sum_{i=1}^{n}(y_{ji} - \bar{y}_j)^2}{n-1} \qquad (11)$$

and an unbiased estimator of $\operatorname{var}(S)$ is identical to (11) after replacing y_{ji} by $S_i = \sum_{j=1}^{k} y_{ji}$ and $\bar{y}_j$ by $\bar{S}$.

In a panel study where S is computed at each wave, let S_w denote the score at wave w and $\hat{\alpha}_w$ the corresponding estimate of α at wave w. In practice, α is estimated separately and independently for each wave. The method of estimating reliability discussed next uses information both within and across waves to assess reliability at each wave.

2.2 Estimating Reliability using the Simplex Model

For panel data, scale score reliability can also be estimated using the so-called simplex model (Heise, 1969; Heise, 1970; Wiley & Wiley, 1970; Jöreskog, 1979). The simplex method uses a longitudinal structural equation model to estimate scale score reliability at each wave using the scale scores themselves (i.e., the S_i's) rather than the responses to the individual items comprising the scale. This is a key advantage of the simplex model over Cronbach's α: since it operates on the aggregate scale scores, correlations between the items within the scale do not bias the estimates of reliability.

To use this method, the same scale must be available from at least three waves of a panel study and the scores must be computed identically at each wave. The covariation of individual scores both within and between the waves provides the basis for an estimate of the reliability of the measurement process. In this sense, the simplex model is akin to a test-retest reliability assessment where the correlation between values of the same variable measured at two or more time points estimates the reliability of those values. An important difference is that while test-retest reliability assumes no change in true score variance or error variance across repeated measurements, the simplex model allows either true score variance to change while holding error variance constant (referred to as the stationary error variance assumption) or vice versa (referred to as the stationary true score variance assumption) according to the situation. Unfortunately, allowing both true score and error variances to vary by wave leads to a non-identified model (i.e., insufficient number of degrees of freedom to obtain a unique solution to the structural equations).

An early version simplex model (proposed by Wiley & Wiley, 1970) assumed stationary error variance and, thus, allowed true score variance to change by wave which seems plausible for most practical situations. In the present work, both types of assumptions (stationary true score variance and stationary error variance) are considered.

This model is composed of a set of measurement equations and structural equations. The measurement equations relate the unobserved true scores to the observed scores.

$$S_w = t_w + \varepsilon_w \tag{12}$$

for w = 1,2,3 where S_w is the observed score, t_w is the unobserved true score (i.e., sum of the k item true scores) and the ε_w is measurement error (i.e., sum of the k item error terms) at wave w=1,2,3.

The structural equations define the relationships among true scores:

$$\begin{aligned} t_2 &= \beta_{12} t_1 + \zeta_2 \\ t_3 &= \beta_{23} t_2 + \zeta_3 \end{aligned} \tag{13}$$

where β_{12} is the effect of the true score at time 1 on the true score at time 2 and β_{23} is the effect of true score at time 2 on true score at time 3. The $\beta_{w,w+1}$ are the parameters that measure change in true score from wave w to wave w+1. The terms ζ_2 and ζ_3 are random error terms that represent the deviations between t_{w+1} and $\beta_{w,w+1} t_w$, sometimes referred to as random shocks. Note that $\operatorname{var}(\zeta_w)$ is a component of true score variance at time w; for example,

$$Var(t_2) = \beta_{12}^2 Var(t_1) + Var(\zeta_2) \tag{14}$$

and

$$\begin{aligned} Var(t_3) &= \beta_{23}^2 Var(t_2) + Var(\zeta_3) \\ &= \beta_{23}^2 \beta_{12}^2 Var(t_1) + \beta_{23}^2 Var(\zeta_2) + Var(\zeta_3) \end{aligned} \tag{15}$$

Assumptions of the simplex model include, for all w, w'=1,2,3

$$\begin{aligned} &E(\varepsilon_w) = 0 \\ &Cov(\varepsilon_w, \varepsilon_{w'}) = 0 \\ &Cov(\varepsilon_w, t_{w'}) = 0 \\ &Cov(\zeta_w, t_{w'}) = 0 \end{aligned} \tag{16}$$

For identification, the original simplex model assumed stationary error variance, that is,

$$Var(\varepsilon_w) = Var(\varepsilon_{w'}) = \sigma_\varepsilon^2 \text{ for } w \neq w' \tag{17}$$

(see Wiley and Wiley, 1970). Stationary true score variance can be substituted for (19) and will be discussed subsequently

The simplex model estimates the parameters β_{12}, β_{23}, σ_ε^2, $\sigma_{t1}^2 = Var(t_1)$, $\sigma_{\zeta 2}^2 = Var(\zeta_2)$, and

$\sigma^2_{\zeta 3} = Var\left(\zeta_3\right)$. The reliabilities for the three waves are given by the following:

$$R_1 = \frac{\sigma^2_{t1}}{\sigma^2_{t1} + \sigma^2_{\varepsilon}} \tag{18}$$

$$R_2 = \frac{\sigma^2_{t2}}{\sigma^2_{t2} + \sigma^2_{\varepsilon}} \tag{19}$$

$$R_3 = \frac{\sigma^2_{t3}}{\sigma^2_{t3} + \sigma^2_{\varepsilon}} \tag{20}$$

Note that equations (18)-(20) all have the same form as (5). If desired, (19) and (20) may be rewritten in terms of β_{12}, β_{23}, σ^2_{ε}, σ^2_{t1}, $\sigma^2_{\zeta 2}$, and $\sigma^2_{\zeta 3}$ using (14) and (15).

Under the Wiley & Wiley simplex model, the error variances are stationary (as in equation 17) and the true score variances are non-stationary. However, there are situations when the error variances should also be non-stationary. For example, the information collected on children for the CBCL may be more subject to random error as the children age. Thus, the error variance at Waves 2 or 3 could be somewhat larger than the error variance at Wave 1. As previously noted, specifying both non-stationary true score and error variances will yield a non-identified model. Thus, if non-stationary error variances are specified, then stationary true score variances must be specified in order to achieve an identified model.

To illustrate, Table 1 provides estimates of reliability for the Youth Self-Report for three waves of the NSCAW. Cronbach's α and the simplex reliability estimates are provided under both the assumptions of stationary error variance and stationary true score variance. The sample sizes varied somewhat for each estimate from 1200 to 1800 cases. Differences as small as 0.05 can be interpreted as significant. Note that the simplex estimates vary considerably within wave: from 0.57 to 0.77 in Wave I. The simplex estimates tend to be smaller than α, substantially so in some cases which suggests that inter-item correlation could be inflating the α estimates of reliability. These results also illustrate the degree to which estimates of R can vary depending upon the method used.

Table 1. YSR Scale Score Reliability Estimates using the Simplex Model and Cronbach's α

Model	Wave 1	Wave 2	Wave 3
Simplex: Stationary Error Variance	0.77	0.71	0.67
Simplex: Stationary True Score Variance	0.57	0.71	0.81
Cronbach's α	0.96	0.95	0.95

Although the simplex model is unaffected by correlated error, it can still be biased due to the failure of other assumptions made in its derivation. As an example, if both error variance and true score variance, the simplex estimates of reliability will be biased regardless of which of these is assumed to be stationary. As an example, suppose that error variances increase over time while true score variance remains constant. In this situation, the reliability ratio actually decreases over time since the denominator increases while the numerator remains constant. The simplex model, under the stationary error variance assumption, will attribute the increase in total variance across time to increasing true score variances. This means that reliability will appear to increase over the time – just the opposite of reality.

The simplex model can also be contaminated to some extent by correlated errors among the waves since it assumes that the score-level errors are independent across time. As an example, if the waves are spaced only a few days apart, subjects may remember their answers from the last interview and repeat them rather than providing independently derived responses. If instead the time interval between waves is a few weeks or more, the risk of recall and consequently between-wave correlated error is much reduced. This may not eliminated the inter-wave correlated error, however. For example, if the subjects tend to misinterpret the items in a scale in the same way at each wave, response errors, even at the aggregate scale-level, could be correlated across waves.

Finally, another assumption of the simplex model is that the ratio of the current wave's true score to the prior wave's true score is a constant apart from the random shock terms. This assumption may not hold in general. For example, some items in the CBCL are specific to a child's age and these items are substituted by other items that are more appropriate for the child as the child ages. Thus, the assumption that the true scores of the scales appropriate to children of all ages satisfy model assumption may be violated and, if so, the simplex model estimates may be unpredictably biased.

The next section introduces a more general model that subsumes the models used to generate the estimates Table 1 as special cases. An important additional feature of the model is that it is identified even if true score and error variances are not stationary; that is, when both are allowed to vary across waves. We also provide an approach for testing which set of model restrictions are satisfied in order to choose the best estimates of reliability.

2.3 The Generalized Simplex Model for Estimating Scale Score Reliability

Using the method of split halves (Brown, 1910; Spearman, 1910), a more general model for estimating scale score reliability can be formulated which relaxes many, but not all, of the assumptions associated with the α and simplex models. Under very general assumptions, this model will provide estimates of reliability for each half of a scale for each wave of data collection. The half-scale reliability estimates for each wave can then be combined to produce a full scale estimate of R_w using a formula similar to the Spearman-Brown Prophecy formula (Carmines & Zeller, 1979) that we have generalized for use when the two half-scales have correlated errors. To simplify the exposition of the model, we assume three panel waves are available; however, extending the model to more than three waves is straightforward.

Suppose the items comprising the score at wave w denoted, S_w, can be split into two equivalent halves. One approach might assign odd numbered items to one half and even number items the other half. However, any method for dividing the items that satisfies the subsequent model assumptions is acceptable. Let S_{w1} and S_{w2} (w = 1,2,3) denote the scores corresponding to the two halves. This model resembles the original simplex model with the only difference being the single score S_w has been replaced by S_{w1} and S_{w2} corresponding to the split halves. Analogous to the simplex model, the generalized (split halves) simplex model assumes the following:

$$\begin{aligned} &\mathrm{E}\left(\varepsilon_{ws}\right) = 0, \text{ for } s = 1,2 \\ &Cov\left(\varepsilon_{ws}, \varepsilon_{w's'}\right) = 0, \text{ for } w \neq w' \qquad (21) \\ &Cov\left(\varepsilon_{ws}, t_{w's'}\right) = 0, \text{ for } (w,s) \neq (w',s') \end{aligned}$$

To be identified, the generalized simplex model requires the restriction that the covariance between the split halves within a wave is constant over time; i.e., $Cov(S_{w1}, S_{w2}) = Cov(S_{w'1}, S_{w'2}) = \sigma_{12}$, say, for all w, w'. We must further assume that the true score variances are equal across the split-halves; that is, $\mathrm{Var}(t_{1w}) = \mathrm{Var}(t_{2w}) = \sigma_{tw}^2$, say. Let $\hat{\sigma}_{tw}^2$ and $\hat{\sigma}_{\varepsilon w}^2$ denote the estimates of the true score and error variances, respectively, for split-halves at wave w and let $\hat{\sigma}_{12}$ denote the estimate of the split-half error covariance at wave w. Then an estimator of the reliability of the score, S_w, is

$$\hat{R}_w = \frac{\hat{\sigma}_{tw}^2}{\hat{\sigma}_{tw}^2 + \hat{\sigma}_{12} + \dfrac{\hat{\sigma}_{\varepsilon w}^2}{2}} \qquad (22)$$

Except for the covariance term in the denominator, this formula is equivalent to the well-known Spearman-Brown prophecy formula (Carmines & Zeller, 1979).

This model can be viewed as a generalization of both α and the simplex models. First, like the simplex model, it is not necessary to assume uncorrelated item-level errors within waves. In addition, the model allows for both non-stationary true score and error variances. Imposing the restriction σ_{12} =0 will produce estimates that are consistent with Cronbach's α. Reliability estimates which are consistent with the simplex model can be produced by specifying either stationary true score variance, error variances or both and removing the constraint $\sigma_{12} = 0$. In this manner, the model can be used in situations where neither α nor the simplex models are appropriate. In these situations, this generalized simplex model will provide better estimates of R_w than either the α or the simplex models. The generalized simplex model can be restricted to test some of the key assumptions of both alternative models: uncorrelated item errors, stationary true score variances and/or stationary error variances.

3. Application: Measures of Child Well-being

In this section, we consider an application of the models in the preceding section for estimating scale score reliability for a number of SSM's obtain in the National Survey of Child Adolescent Well-being (NSCAW). The NSCAW is a panel survey of about 5100 children who were investigated for child abuse or neglect in 87 randomly selected U.S. counties (Dowd, et al, 2004). An important component of the data quality evaluation for this survey was the assessment of reliability for all the key SSM's. Biemer, et al (2006) provided estimates for more than 30 SSM's using both Cronbach's α and the simplex model assuming stationary true score variances, stationary error variances or both. A representative subset of these scores was considered including: the Child Behavior Checklist (CBCL), Teacher Report Form (TRF), the Youth Self-Report (YSR) and the Short-Form Health Survey (SF-12). Table 2 presents the reliability estimates and their standard errors for the total CBCL score and seven models

Similar to the total CBCL estimates, for all of the other SSM's, the simplex model estimates are lower than the Cronbach α estimates. For cases where the simplex variance stationarity assumptions matter, including the total CBCL, reliability estimates tend to decrease over time for the original simplex model while the opposite is true for the simplex estimates that assume stationary true score variance. To understand why this makes sense, recall that, under our models, total variance is the sum of true score and error variance. If true score variance constrained in the model, then any change in true score variance across time will be attributed to a change in error variance. Since an increase in error variance will decrease reliability (assuming the true score variance is constant), reliability will appear to increase under this constraint. Likewise, if error variance is constrained, then changes in the error variance across time will be attributed to changes in the true score variance. Since an increase in true score variance will increase reliability if the error variance is held constant, reliability will appear to increase under this constraint. Thus, the two assumptions will produce opposing effects on the reliability.

Estimates obtained from the generalized models with either stationary true score or stationary error variance constraints are very close to the simplex models with these same constraints. The magnitude of the reliability estimates is comparable to the original simplex model estimates for many measures, but is generally lower for the SF-12 measures. For almost all measures the generalized simplex model with uncorrelated errors produces higher reliability estimates than the generalized model with correlated errors. The latter estimates are in close agreement with the α estimates. Estimates from the generalized simplex model without constraints are most similar to the generalized original simplex model. In fact, the estimates at wave 3 are the same or nearly the same for both models. This suggests that the original simplex model may be preferred over both α and the simplex model with stationary true score variance constraints in most practical situations

Table 2: Reliability Estimates for CBCL (2+ years) Total Problem Behavior

N (by wave for α)	Model	Reliability (standard errors)		
		Wave 1	Wave 3	Wave 4
5330	Simplex Model (SM)	0.756 (0.019)	0.732 (0.020)	0.725 (0.023)
	SM Stationary True Score Variance	0.666 (0.030)	0.732 (0.051)	0.753 (0.122)
589, 985, 1259	Coefficient α (2-3 years) (98 items)	0.942 (0.004)	0.945 (0.003)	0.949 (0.002)
3174, 3002, 3359	Coefficient α (4+ years) (118 items)	0.962[1]	0.962[1]	0.962[1]
5330	Generalized Model (GM)	0.650 (0.022)	0.612 (0.022)	0.600 (0.025)
	GM Stationary True Score Variance	0.718 (0.017)	0.707 (0.016)	0.709 (0.016)
	GM Stationary Error Variance	0.647 (0.022)	0.615 (0.022)	0.602 (0.025)
	GM Uncorrelated Error	0.898 (0.004)	0.875 (0.004)	0.874 (0.003)

[1] Models run in SAS proc mixed. No standard error estimates available.

Tests of the stationary error variances, stationary true score variance, and uncorrelated error assumptions can be done in the context of the generalized simplex model where the models with constraints are nested in the larger generalized model without constraints. Results from these tests for all measures are described subsequently. Results for the three CBCL measures are presented in Table 3. The assumption of uncorrelated errors was rejected for all 11 SSM's considered. The assumption of stationary error variance was rejected for eight SSM's as was the assumption of stationary true score variance. In seven cases, neither of these assumptions could be accepted.

These results suggest that α is not an appropriate indicator of reliability all 11 SSM's considered in our study. The two simplex models performed much better in terms of producing estimates of R that are close to those of the generalized simplex model. However, the stationary variance assumptions were often rejected. If one had to choose, the original simplex model estimates seemed to agree more often and closely with the estimates from the generalized simplex model.

Table 3: Nested Wald Tests for the CBCL SSM's

Measure	N	Uncorrelated Errors			Stationary Error Variance			Stationary True Score Variance		
		Chi-square	DF	p-value	Chi-square	DF	p-value	Chi-square	DF	p-value
CBCL (2+ years) Total Problem Behavior	5330	122.786	1	0.000	7.613	2	0.0222	16.692	2	0.0002
CBCL (2+ years) Externalizing	5330	78.244	1	0.000	3.650	2	0.1612	18.660	2	0.0001
CBCL (2+ years)	5330	65.005		0.000	28.805		0.0000	15.411		

Internalizing	1	2	2	0.0005

4. Conclusions

This analysis suggests that the choice of model and assumptions is critical in the evaluation of scale score reliability. Blind use of Cronbach's α can and often does lead to a biased assessment of the reliability of SSM's. In our study, assumption of inter-item uncorrelated error, upon which α relies, was rejected for all the SSM's we considered. Consequently, $\hat{\alpha}$ was higher, often exceedingly so, than the simplex estimates which do not require that assumption. When panel data are available, the simplex model with either the stationary error or true score variance assumption can be employed and will permit a more valid assessments of reliability. However, as we have shown in Table 3, the assumptions underlying the simplex approach also do not hold for many SSM's. In such cases, more valid estimates of reliability can be obtained using the generalized simplex model which requires neither the variance stationarity nor uncorrelated inter-item error assumptions to provide valid estimates of reliability.

References:

Achenbach, T. M. (1991a). Manual for the Child Behavior Checklist 2 - 3 and 1991 profile. Burlington, Department of Psychiatry, University of Vermont.

Achenbach, T. M. (1991b). Manual for the Child Behavior Checklist 4 - 18 and 1991 profile. Burlington, Department of Psychiatry, University of Vermont.

Biemer, P. P., Christ, S. L., & Wiesen, C. A. (2006). Scale score reliability in the National Survey of Child and Adolescent Well-being. Internal Report. RTP, NC: RTI International.

Biemer, P.P., & D. Trewin (1997). A review of measurement error effects on the analysis of survey data. In Lyberg, L. et al. (Eds.), Survey measurement and process quality. New York: John Wiley & Sons, pp. 603-632.

Bollen, Kenneth A. (1989). Structural equations with latent variables. John Wiley & Sons, New York.

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

Carmines, E. G. & Zeller, R. A. (1979). Reliability and Validity Assessment. In E. G. Carmines (Ed.) Sage university papers series on quantitative applications in the social sciences. 107-117. Newbury Park, CA: Sage.

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, 297-334.

Dowd, K., S. Kinsey, S. Wheeless, S. Suresh & the NSCAW Research Group (2004). National Survey of Child and Adolescent Well-being: Combined Waves 1-4 data file user's manual. Research Triangle Park, NC: RTI International

Fuller, W. A. (1987). Measurment error models. New York: Wiley & Sons.

Green, S.B. and Hershberger, S.L. (2000). Correlated errors in true score models and their effect on Coefficient Alpha. Structural equation modeling, 7, 251-270.

Heise, D.R. (1970). Separating reliability and stability in test-retest correlation. American sociological review, 34, 93-101.

Heise, D.R. (1969). Comment on "The estimation of measurement error in panel data". American sociological review, 35, 117.

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.

Jöreskog, Karl G. (1979). Statistical models and methods for analysis of longitudinal data. In Advances in factor analysis and structural equation models. Jöreskog & Sörbom, Eds. Cambridge, MA: Abt Books.

Komaroff, E. (1997). Effect of simultaneous violations of essential τ -equivalence and uncorrelated error on coefficient α. Applied psychological measurement, 21: 337 - 348.

Lucke, J.E. (2005). "Rassling the hog": The influence of correlated item error on internal consistency, classical reliability, and congeneric reliability. Applied psychological measurement,. 29, 106-125.

Rae, G. (2006). Correcting Coefficient Alpha for correlated errors: Is α_K a lower bound to reliability? Applied psychological measurement, 30, 56-59.

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

Raykov, T. (2001). Bias of coefficient alpha for fixed congeneric measures with correlated errors. Applied psychological measurement, 25, 69-76.

Raykov, T., Shrout, P.E. (2002). Reliability of scales with general structure: point and interval estimation using a structural equation modeling approach. Structural equation modeling. 9(2): 195-212.

Spearman, C. (1910) Correlation calculated from faulty data. British journal of psychology, 3, 271-295.

Vermunt, J. (1996). Log-linear models for event histories, Sage Publications, Thousand Oaks, CA.