A General Approach for Estimating Scale-Score Reliability for Panel Data

Paul Biemer, Sharon Christ and Christopher Wiesen

Odum Institute for Research in Social Science

University of North Carolina at Chapel Hill

Abstract

Scale score measures are ubiquitous in the psychological literature and can be used as both dependent and independent variables in data analysis. Poor reliability of scale score measures leads to inflated standard errors and/or analytic invalidity. Cronbach's α is often employed to assess the reliability but, due to its rather strong assumptions, can be a poor indicator. For panel data, an alternative approach is the simplex method; however, it too requires assumptions that may not hold in practice. In this paper, a new estimator of reliability is proposed which relaxes the assumptions of both Cronbach's α and the simplex estimator and, in that sense, is a generalization of both estimators. The virtues of new method are illustrated using data from a large scale panel survey.

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, as reliability decreases, the standard errors of estimates of means, totals and proportions increase. 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 alpha with relative ease. There are numerous examples in the literature of using α for assessing the reliability of scale scores (see, for example, Burney & Kromrey, 2001; Sapin, et al., 2005; Yoshizumi, Murakami, & Takai, 2006). 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 severely 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). This assumption 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 measure anger or pain instead. 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[1] data, there are alternatives to alpha that rely on assumptions that are more easily satisfied in practice. One of these is the *simplex* estimator of reliability (Wiley and Wiley, 1970).

---

[1] A panel study collects data from the same subjects (i.e., a panel) at different points in time, usually at regular intervals.

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 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 *i*th 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 *i*th 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

$$y_i = \mu_i + b_i + \varepsilon_i$$

or equivalently,

$$y_i = t_i + \varepsilon_i$$

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

$$\sigma_{\varepsilon i}^2 = \mathrm{E}(\varepsilon_i^2 \mid i)$$

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

y_i is given by

$$\text{Var}(y_i) = \text{Var}(t_i) + \text{Var}(\varepsilon_i)$$
$$= \sigma_t^2 + \sigma_\varepsilon^2$$

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

$$R = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_\varepsilon^2}$$

(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. 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 alpha 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 Figure 1 and corresponding model equations as follows:

$$y_1 = \lambda_1 t + \varepsilon_1$$
$$y_2 = \lambda_2 t + \varepsilon_2$$
$$y_3 = \lambda_3 t + \varepsilon_3$$
$$y_4 = \lambda_4 t + \varepsilon_4$$

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.

[INSERT FIGURE 1 ABOUT HERE]

The model in Figure 1 also assumes that the measurement errors, ε_j's are uncorrelated between items; i.e., $\mathrm{cov}\left(\varepsilon_j, \varepsilon_{j'}\right) = 0$ for any two items $j \text{ and } j'$. This assumption is indicated by the lack of arrows between the ε_j's in the figure. In addition, Cronbach's α assumes that the four measurements are *parallel*; that is, $\lambda_j = 1$ and $Var\left(\varepsilon_j\right) = \sigma_\varepsilon^2$, 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

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

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

$$R = \frac{k^2 \sigma_t^2}{k^2 \sigma_t^2 + k \sigma_\varepsilon^2}$$

or

$$R = \frac{\sigma_t^2}{\sigma_t^2 + \dfrac{\sigma_\varepsilon^2}{k}}$$

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

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

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

$$\mathrm{var}(y_j) = \frac{\sum\limits_{i=1}^{n} (y_{ji} - \overline{y}_j)^2}{n-1}$$

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

In a panel survey where S is computed at each wave, let S_w denote the score at wave w and $\hat{\alpha}_w$ the corresponding estimate of alpha 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 survey 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).

One early version of the simplex model (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.

The original simplex model for three repeated measurements is illustrated in Figure 2.

[INSERT FIGURE 2 ABOUT HERE]

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.

8

$$S_w = t_w + \varepsilon_w$$

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. From Figure 2, we see that

$$t_2 = \beta_{12}t_1 + \zeta_2$$
$$t_3 = \beta_{23}t_2 + \zeta_3$$

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 $\mathrm{var}(\zeta_w)$ is a component of true score variance at time w; for example,

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

and

$$Var(t_3) = \beta_{23}^2 Var(t_2) + Var(\zeta_3)$$
$$= \beta_{23}^2\beta_{12}^2 Var(t_1) + \beta_{23}^2 Var(\zeta_2) + Var(\zeta_3)$$

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

$$E(\varepsilon_w) = 0$$
$$Cov(\varepsilon_w, \varepsilon_{w'}) = 0$$
$$Cov(\varepsilon_w, t_{w'}) = 0$$
$$Cov(\zeta_w, t_{w'}) = 0$$

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

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

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

$$R_1 = \frac{\sigma_{t1}^2}{\sigma_{t1}^2 + \sigma_\varepsilon^2}$$

$$R_2 = \frac{\sigma_{t2}^2}{\sigma_{t2}^2 + \sigma_\varepsilon^2}$$

$$(10)$$

$$R_3 = \frac{\sigma_{t3}^2}{\sigma_{t3}^2 + \sigma_\varepsilon^2}$$

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_ε^2, σ_{t1}^2, $\sigma_{\zeta 2}^2$, and $\sigma_{\zeta 3}^2$ using (14) and (15).

Under the Wiley & Wiley simplex model, the error variances are stationary (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[2].

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

---

[2] It is also possible to obtain an identified model assuming the reliability ratio is constant over waves (i.e., stationary reliability). The case was considered in our work but not reported here to save space. This produced reliability estimates that were constant across waves and approximately equal to the average reliability obtained by the alternative stationarity assumptions.

reliability. These results also illustrate the degree to which estimates of *R* can vary depending upon the method used.

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.

[TABLE 1 ABOUT HERE]

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 (see Figure 2). 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. The path model summarizing the assumptions for the split halves model is shown in Figure 3. Note its resemblance to the model in Figure 2 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:

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

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

12

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

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

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 $\sigma_{12} = 0$ 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.

[INSERT FIGURE 3 ABOUT HERE]

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 will be considered here including: the Child Behavior Checklist (CBCL), Teacher Report Form (TRF),

Self-Report Instrument for Middle School Students (RAPS-SM), the Youth Self-Report (YSR) and the Short-Form Health Survey (SF-12).

Table 2 presents the reliability estimates and their standard errors for 19 SSM's and seven models:

1. simplex model with stationary error variance (referred to as the *original simplex model*),
2. simplex model with stationary true score variance,
3. Cronbach's α,
4. generalized simplex model without stationarity constraints,
5. generalized model with stationary true score variance,
6. generalized model with stationary error variance, and
7. generalized model with uncorrelated errors and without stationarity constraints.

Almost universally, the original simplex model estimates are lower than the Cronbach α estimates. There are a few exceptions, specifically some of the RAPS scores, where the estimates are not markedly different across these models. Noted that for cases where the simplex estimate assumptions matter, reliability estimates tend to decrease over time for the original simplex model while the opposite is true for the simplex estimates assuming 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 (assuming the error variance is constant), reliability will appear to increase under this constraint. Thus, the two assumptions will produce opposing effects on the reliability.

[Table 2 about here]

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 many of the RAPS measures and 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

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 are presented in Table 3. For only 2 out of 19 SSM's – specifically the RAPS Autonomy Support measures – could the assumption of uncorrelated errors not be rejected. For those measures, the reliability estimates for the models with and without uncorrelated errors were very similar (Table 2). For seven SSM's, the assumption of constant error variance could not be rejected and for ten measures the assumption of constant true score variance could not be rejected.

These results suggest that α is not an appropriate indicator of reliability for all but 2 SSM's considered in our study. For neither of the simplex models was performance exemplary. Both suffered some bias owing ostensibly to violations of the stationarity assumption. 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 about here]

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 an over-optimistic assessment of the reliability of SSM's. When the data allow it, employing the original Wiley and Wiley simplex model will lead to more valid assessments of reliability. However, as we have shown in Table 3, the assumptions underlying this 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.

One limitation of the generalized simplex model is its reliance on split half scores. For SSM's having a small number of items (say, less than 10), it may not be possible to form

equivalent split halves in which case the assumptions of the model would be violated. In such situations, we recommend the original simplex model be used. Alternatively, the generalized model could be applied for more than one split of the items in order to gauge the degree to which the estimates depend on the particular split used in the analysis. This practice was not followed in the present study.

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., and 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., and 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*. Wiley Series in Probability and Mathematical Statistics. New York: Wiley.

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

Burney, D. M., and Kromrey, J. (2001). Initial development and score validation of the Adolescent Anger Rating Scale. *Educational and psychological measurement, 61*, 446-460.

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). <u>Measurment Error Models</u>. 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., and 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 <u>Advances in Factor Analysis and Structural Equation Models</u>. 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.

Sapin, C., Simeoni, M. C., El Khammar, M., Antoniotti, S. Auquier, P. (2005). Reliability and validity of the VSP-A, a health-related quality of life instrument for ill and healthy adolescents. *Journal of Adolescent Health, 36,* 327-336.

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.

Vehkalahti, K., Puntanen, S., and Tarkkonen, L. (2006). *Estimation of reliability: a Better alternative for Cronbach's Alpha*. Reports on Mathematics, Preprint 430, Department of Mathematics and Statistics, University of Helsinki, Finland.

Wiley, D. E. & Wiley, J.A.. (1970). The Estimation of Measurement Error in Panel Data. American Sociological Review, *35*, 112-117.

Yoshizumi, T., Murase, S., Murakami, T., & Takai, J. (2006). Reliability and validity of the Parenting Scale of Inconsistency. *Psychological reports, 99*, 74-84.

Zimmerman, D. W., & Zumbo, B. D. (1993). Coefficient Alpha as an estimate of test reliability under violation of two assumptions. *Educational & psychological measurement, 53*, 33-50.

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

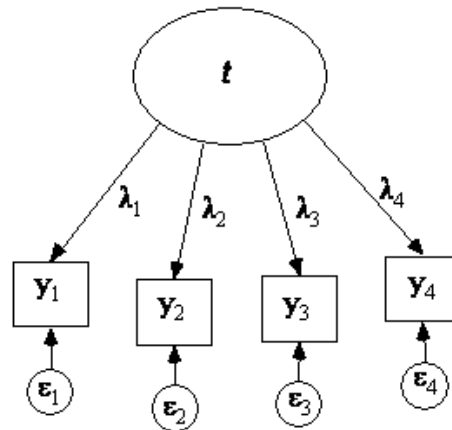| MODEL	WAVE 1	WAVE 2	WAVE 3
Simplex Model			
Constant Error Variance	0.77	0.71	0.67
Constant True Score Variance	0.57	0.71	0.81
Cronbach's Alpha	0.96	0.95	0.95

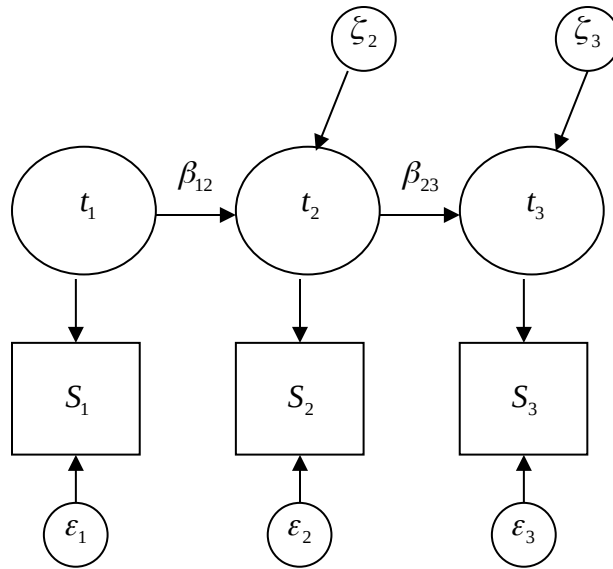**Figure 1. Cronbach's Alpha Model for Four Indicators**

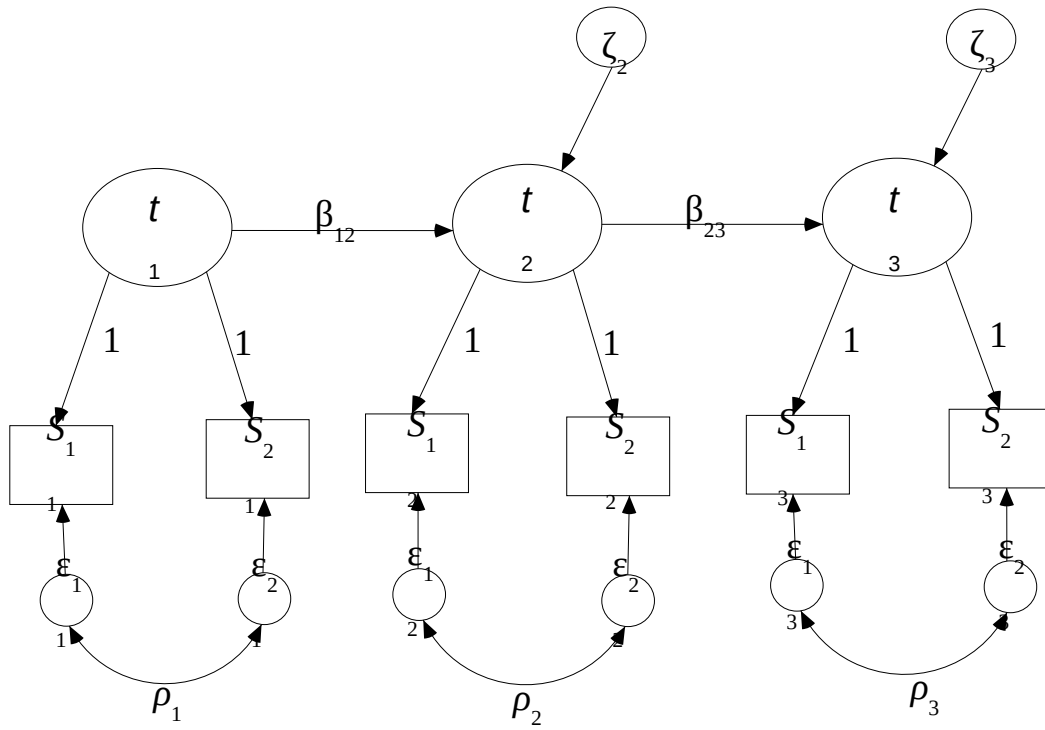**Figure 2. Simplex Model for Three Repeated Scores**

Figure 3. Generalized Simplex Model for Three Repeated Scores

Table 2: Reliability Estimates for Select NSCAW Measures

Measure	*N* (by wave for alpha)	Model	Reliability (standard errors)		
			Wave 1	**Wave 3**	**Wave 4**
CBCL (2+ years) Total Problem Behavior	5330	Simplex Model (SM)	0.756 (0.019)	0.732 (0.020)	0.725 (0.023)
		SM Constant True Score Variance	0.666 (0.030)	0.732 (0.051)	0.753 (0.122)
	589, 985, 1259	Coefficient Alpha (2-3 years) (98 items)	0.942 (0.004)	0.945 (0.003)	0.949 (0.002)
	3174, 3002, 3359	Coefficient Alpha (4+ years) (118 items)	0.962[2]	0.962[2]	0.962[2]
	5330	Generalized Model (GM)	0.650 (0.022)	0.612 (0.022)	0.600 (0.025)
		GM Constant True Score Variance	0.718 (0.017)	0.707 (0.016)	0.709 (0.016)
		GM Constant 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)
CBCL (2+ years) Externalizing	5330	Simplex Model (SM)	0.763 (0.021)	0.737 (0.020)	0.731 (0.023)
		Constant True Score Variance	0.667 (0.028)	0.738 (0.051)	0.757 (0.111)
	589, 985, 1259	Coefficient Alpha (2-3 years) (26 items)	0.906 (0.006)	0.903 (0.005)	0.909 (0.004)
	3174, 3002, 3359	Coefficient Alpha (4+ years) (33 items)	0.928 (0.002)	0.929 (0.002)	0.928 (0.002)
	5330	Generalized Model (GM)	0.768 (0.020)	0.747 (0.020)	0.739 (0.023)
		GM Constant True Score Variance	0.823 (0.013)	0.827 (0.013)	0.828 (0.013)
		GM Constant Error Variance	0.772 (0.020)	0.747 (0.020)	0.738 (0.022)
		GM Uncorrelated Error	0.935 (0.002)	0.932 (0.003)	0.931 (0.002)
CBCL (2+ years) Internalizing	5330	Simplex Model (SM)	0.714 (0.026)	0.677 (0.029)	0.656 (0.033)
		Constant True Score Variance	0.598 (0.039)	0.677 (0.069)	0.720 (0.149)
	589, 985, 1259	Coefficient Alpha (2-3 years) (25 items)	0.818 (0.013)	0.833 (0.009)	0.844 (0.007)
	3174, 3002, 3359	Coefficient Alpha (4+ years) (31 items)	0.900 (0.003)	0.900 (0.003)	0.895 (0.003)

Measure	N (by wave for alpha)	Model	Reliability (standard errors)		
			Wave 1	**Wave 3**	**Wave 4**
CBCL (2+ years) Internalizing	5330	Generalized Model (GM)	0.692 (0.024)	0.667 (0.027)	0.647 (0.030)
		GM Constant True Score Variance	0.739 (0.018)	0.751 (0.018)	0.759 (0.018)
		GM Constant Error Variance	0.702 (0.024)	0.667 (0.027)	0.639 (0.030)
		GM Uncorrelated Error	0.887 (0.004)	0.883 (0.004)	0.885 (0.004)
YSR (11+ years) Total Problem Behavior	1832	Simplex Model (SM)	0.768 (0.047)	0.710 (0.054)	0.669 (0.069)
		Constant True Score Variance	0.569 (0.048)	0.710 (0.133)	0.812 (0.259)
	1167, 1270, 1590	Coefficient Alpha (101 items)	0.956 (0.003)	0.950 (0.002)	0.945 (0.002)
	1825	Generalized Model (GM)	0.750 (0.048)	0.692 (0.053)	0.662 (0.065)
		GM Constant True Score Variance	0.793 (0.037)	0.799 (0.038)	0.806 (0.038)
		GM Constant Error Variance	0.755 (0.048)	0.692 (0.053)	0.655 (0.065)
		GM Uncorrelated Error	0.953 (0.003)	0.946 (0.003)	0.948 (0.003)
YSR (11+ years) Externalizing	1832	Simplex Model (SM)	0.788 (0.039)	0.733 (0.047)	0.721 (0.054)
		Constant True Score Variance	0.583 (0.045)	0.733 (0.118)	0.766 (0.216)
	1167, 1270, 1590	Coefficient Alpha (30 items)	0.895 (0.006)	0.876 (0.005)	0.877 (0.005)
	1825	Generalized Model (GM)	0.747 (0.043)	0.691 (0.047)	0.700 (0.052)
		GM Constant True Score Variance	0.751 (0.034)	0.761 (0.032)	0.775 (0.032)
		GM Constant Error Variance	0.753 (0.041)	0.691 (0.047)	0.682 (0.052)
		GM Uncorrelated Error	0.877 (0.009)	0.850 (0.009)	0.867 (0.008)

Measure	*N* (by wave for alpha)	Model	Reliability (standard errors)		
			Wave 1	**Wave 3**	**Wave 4**
YSR (11+ years) Internalizing	1832	Simplex Model (SM)	0.717 (0.068)	0.668 (0.073)	0.595 (0.103)
		Constant True Score Variance	0.569 (0.067)	0.668 (0.176)	0.815 (0.351)
	1167, 1270, 1589	Coefficient Alpha (31 items)	0.899 (0.006)	0.894 (0.006)	0.879 (0.006)
	1825	Generalized Model (GM)	0.708 (0.067)	0.656 (0.071)	0.602 (0.093)
		GM Constant True Score Variance	0.736 (0.052)	0.743 (0.052)	0.757 (0.053)
		GM Constant Error Variance	0.712 (0.066)	0.656 (0.071)	0.584 (0.092)
		GM Uncorrelated Error	0.900 (0.007)	0.884 (0.009)	0.879 (0.007)
TRF (5+ years) Total Problem Behavior	2631	Simplex Model (SM)	0.604 (0.074)	0.589 (0.071)	0.552 (0.089)
		Constant True Score Variance	0.566 (0.073)	0.589 (0.156)	0.642 (0.258)
	1325, 1394, 1610	Coefficient Alpha (95 items)	0.966 (0.001)	0.967 (0.001)	0.967 (0.001)
	2643	Generalized Model (GM)	0.599 (0.078)	0.582 (0.075)	0.546 (0.085)
		GM Constant True Score Variance	0.717 (0.061)	0.720 (0.061)	0.720 (0.061)
		GM Constant Error Variance	0.601 (0.078)	0.582 (0.075)	0.546 (0.086)
		GM Uncorrelated Error	0.969 (0.002)	0.970 (0.002)	0.968 (0.002)
TRF (5+ years) Externalizing	2635	Simplex Model (SM)	0.656 (0.065)	0.640 (0.064)	0.602 (0.076)
		Constant True Score Variance	0.612 (0.068)	0.640 (0.145)	0.707 (0.269)
	1325, 1393, 1610	Coefficient Alpha (28 items)	0.952 (0.002)	0.952 (0.002)	0.950 (0.002)
	2642	Generalized Model (GM)	0.651 (0.064)	0.638 (0.063)	0.605 (0.070)
		GM Constant True Score Variance	0.746 (0.047)	0.750 (0.047)	0.754 (0.047)
		GM Constant Error Variance	0.655 (0.065)	0.638 (0.063)	0.601 (0.071)
		GM Uncorrelated Error	0.944 (0.003)	0.946 (0.003)	0.945 (0.003)

Measure	N (by wave for alpha)	Model	Reliability (standard errors)		
			Wave 1	**Wave 3**	**Wave 4**
TRF (5+ years) Internalizing	2521	Simplex Model (SM)	0.389 (0.091)	0.333 (0.090)	0.283 (0.125)
		Constant True Score Variance	0.306 (0.087)	0.333 (0.152)	0.358 (0.210)
	1323, 1394, 1609	Coefficient Alpha (31 items)	0.913 (0.004)	0.910 (0.004)	0.906 (0.005)
	2642	Generalized Model (GM)	0.429 (0.091)	0.368 (0.094)	0.331 (0.119)
		GM Constant True Score Variance	0.515 (0.099)	0.514 (0.098)	0.527 (0.099)
		GM Constant Error Variance	0.427 (0.090)	0.367 (0.093)	0.311 (0.120)
		GM Uncorrelated Error	0.926 (0.005)	0.917 (0.005)	0.929 (0.005)
RAPS (11+ years) Emotional Security Primary Caregiver	1821	Simplex Model (SM)	0.643 (0.110)	0.644 (0.108)	0.608 (0.129)
		Constant True Score Variance	0.645 (0.115)	0.644 (0.246)	0.708 (0.412)
	1159, 1267, 1562	Coefficient Alpha (3 items)	0.649 (0.024)	0.682 (0.021)	0.697 (0.018)
	1821	Generalized Model (GM)	0.362 (0.086)	0.398 (0.094)	0.403 (0.102)
		GM Constant True Score Variance	0.400 (0.058)	0.430 (0.063)	0.451 (0.067)
		GM Constant Error Variance	0.400 (0.089)	0.383 (0.088)	0.355 (0.089)
		GM Uncorrelated Error	0.449 (0.029)	0.493 (0.029)	0.524 (0.023)
RAPS (11+ years) Emotional Security Secondary Caregiver	1134	Simplex Model (SM)	0.703 (0.130)	0.613 (0.162)	0.649 (0.162)
		Constant True Score Variance	0.472 (0.126)	0.613 (0.360)	0.556 (0.430)
	579, 610, 751	Coefficient Alpha (3 items)	0.776 (0.026)	0.769 (0.024)	0.785 (0.023)
	1134	Generalized Model (GM)	0.420 (0.130)	0.362 (0.136)	0.399 (0.141)
		GM Constant True Score Variance	0.433 (0.106)	0.483 (0.110)	0.483 (0.111)
		GM Constant Error Variance	0.465 (0.122)	0.348 (0.128)	0.384 (0.132)
		GM Uncorrelated Error	0.622 (0.036)	0.624 (0.034)	0.660 (0.028)

Measure	N (by wave for alpha)	Model	Reliability (standard errors)		
			Wave 1	Wave 3	Wave 4
RAPS (11+ years) Involvement Primary Caregiver	1821	Simplex Model (SM)	0.490 (0.069)	0.486 (0.073)	0.466 (0.083)
		Constant True Score Variance	0.483 (0.085)	0.486 (0.137)	0.505 (0.216)
	1155, 1266, 1561	Coefficient Alpha (4 items)	0.624 (0.021)	0.607 (0.020)	0.573 (0.019)
RAPS (11+ years) Involvement Primary Caregiver	1821	Generalized Model (GM)	0.190 (0.046)	0.185 (0.047)	0.172 (0.045)
		GM Constant True Score Variance	0.215 (0.035)	0.214 (0.035)	0.213 (0.035)
		GM Constant Error Variance	0.188 (0.045)	0.185 (0.047)	0.173 (0.045)
		GM Uncorrelated Error	0.282 (0.018)	0.273 (0.018)	0.253 (0.013)
RAPS (11+ years) Involvement Secondary Caregiver	1133	Simplex Model (SM)	0.607 (0.135)	0.602 (0.133)	0.600 (0.148)
		Constant True Score Variance	0.593 (0.139)	0.602 (0.294)	0.603 (0.406)
	580, 610, 750	Coefficient Alpha (4 items)	0.672 (0.027)	0.701 (0.026)	0.682 (0.021)
	1133	Generalized Model (GM)	0.259 (0.105)	0.262 (0.105)	0.257 (0.109)
		GM Constant True Score Variance	0.293 (0.066)	0.304 (0.069)	0.301 (0.070)
		GM Constant Error Variance	0.273 (0.111)	0.262 (0.105)	0.262 (0.108)
		GM Uncorrelated Error	0.360 (0.027)	0.373 (0.029)	0.353 (0.021)
RAPS (11+ years) Autonomy Support Primary Caregiver	1817	Simplex Model (SM)	0.433 (0.081)	0.392 (0.083)	0.397 (0.096)
		Constant True Score Variance	0.366 (0.080)	0.392 (0.146)	0.389 (0.190)
	1148, 1261, 1558	Coefficient Alpha (2 items)	0.357 (0.036)	0.354 (0.039)	0.333 (0.040)
	1817	Generalized Model (GM)	0.409 (0.094)	0.411 (0.085)	0.391 (0.106)
		GM Constant True Score Variance	0.378 (0.068)	0.399 (0.067)	0.396 (0.070)
		GM Constant Error Variance	0.452 (0.082)	0.410 (0.085)	0.406 (0.089)
		GM Uncorrelated Error	0.357 (0.036)	0.363 (0.033)	0.335 (0.040)

Measure	N (by wave for alpha)	Model	Reliability (standard errors)		
			Wave 1	Wave 3	Wave 4
RAPS (11+ years) Autonomy Support Secondary Caregiver	1130	Simplex Model (SM)	0.411 (0.104)	0.320 (0.106)	0.344 (0.127)
		Constant True Score Variance	0.277 (0.092)	0.320 (0.175)	0.309 (0.201)
	577, 610, 749	Coefficient Alpha (2 items)	0.358 (0.052)	0.320 (0.067)	0.283 (0.058)
	1130	Generalized Model (GM)	0.342 (0.115)	0.298 (0.099)	0.264 (0.124)
		GM Constant True Score Variance	0.280 (0.081)	0.304 (0.087)	0.291 (0.084)
		GM Constant Error Variance	0.391 (0.098)	0.298 (0.099)	0.319 (0.109)
		GM Uncorrelated Error	0.358 (0.052)	0.315 (0.063)	0.283 (0.058)
RAPS (11+ years) Structure Primary Caregiver	1817	Simplex Model (SM)	0.513 (0.083)	0.499 (0.088)	0.482 (0.105)
		Constant True Score Variance	0.484 (0.095)	0.498 (0.172)	0.515 (0.236)
	1148, 1258, 1556	Coefficient Alpha (3 items)	0.524 (0.030)	0.555 (0.027)	0.547 (0.020)
	1817	Generalized Model (GM)	0.247 (0.060)	0.261 (0.057)	0.250 (0.066)
		GM Constant True Score Variance	0.267 (0.044)	0.286 (0.044)	0.292 (0.047)
		GM Constant Error Variance	0.285 (0.058)	0.262 (0.057)	0.242 (0.061)
		GM Uncorrelated Error	0.308 (0.027)	0.340 (0.025)	0.333 (0.019)
RAPS (11+ years) Structure Secondary Caregiver	1131	Simplex Model (SM)	0.703 (0.211)	0.663 (0.243)	0.669 (0.255)
		Constant True Score Variance	0.584 (0.225)	0.663 (0.559)	0.651 (0.767)
	579, 610, 749	Coefficient Alpha (3 items)	0.579 (0.032)	0.592 (0.035)	0.622 (0.027)
	1131	Generalized Model (GM)	0.323 (0.123)	0.322 (0.134)	0.350 (0.139)
		GM Constant True Score Variance	0.326 (0.090)	0.355 (0.096)	0.370 (0.099)
		GM Constant Error Variance	0.371 (0.129)	0.323 (0.133)	0.322 (0.134)
		GM Uncorrelated Error	0.377 (0.032)	0.387 (0.034)	0.412 (0.030)

Measure	*N* (by wave for alpha)	Model	Reliability (standard errors)		
			Wave 1	Wave 3	Wave 4
SF-12 Mental Health of Caregiver	5484	Simplex Model (SM)	0.532 (0.029)	0.501 (0.030)	0.496 (0.035)
		Constant True Score Variance	0.470 (0.033)	0.501 (0.058)	0.506 (0.098)
	5387, 4652, 4615	Coefficient Alpha (12 items)	0.621 (0.006)	0.616 (0.008)	0.614 (0.008)
	5491	Generalized Model (GM)	0.420 (0.030)	0.397 (0.028)	0.389 (0.033)
		GM Constant True Score Variance	0.454 (0.026)	0.471 (0.026)	0.469 (0.027)
		GM Constant Error Variance	0.438 (0.029)	0.395 (0.028)	0.390 (0.032)
		GM Uncorrelated Error	0.696 (0.009)	0.691 (0.011)	0.692 (0.011)
SF-12 Physical Health of Caregiver	5484	Simplex Model (SM)	0.578 (0.028)	0.587 (0.026)	0.616 (0.027)
		Constant True Score Variance	0.600 (0.029)	0.587 (0.056)	0.546 (0.073)
	5387, 4652, 4615	Coefficient Alpha (12 items)	0.676 (0.006)	0.689 (0.006)	0.696 (0.006)
	5491	Generalized Model (GM)	0.241 (0.030)	0.268 (0.022)	0.282 (0.027)
		GM Constant True Score Variance	0.289 (0.021)	0.294 (0.021)	0.283 (0.020)
		GM Constant Error Variance	0.249 (0.029)	0.265 (0.022)	0.294 (0.027)
		GM Uncorrelated Error	0.413 (0.015)	0.411 (0.017)	0.441 (0.015)

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

Table 3: Nested Wald Tests for the SSM's in Table 2*

Measure	N	Uncorrelated Errors			Constant Error Variance			Constant 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) Internalizing	5330	65.005	1	0.000	28.805	2	0.0000	15.411	2	0.0005
YSR (11+ years) Total Problem Behavior	1825	21.031	1	0.000	18.360	2	0.0001	23.782	2	0.000
YSR (11+ years) Externalizing	1825	10.681	1	0.0011	10.629	2	0.0049	18.613	2	0.0001
YSR (11+ years) Internalizing	1825	9.583	1	0.0020	7.307	2	0.0259	25.084	2	0.000
TRF (5+ years) Total Problem Behavior	2643	25.338	1	0.000	**2.135**	**2**	**0.3438**	**4.688**	**2**	**0.0959**
TRF (5+ years) Externalizing	2642	24.243	1	0.000	6.415	2	0.0405	6.042	2	0.0488
TRF (5+ years) Internalizing	2642	31.154	1	0.000	15.481	2	0.0004	**5.494**	**2**	**0.0641**
RAPS (11+ years) Emotional Security Primary Caregiver	1821	184.206	3	0.000	10.974	2	0.0041	**0.295**	**2**	**0.8630**
RAPS (11+ years) Emotional Security Secondary Caregiver	1134	89.607	3	0.000	6.249	2	0.0440	5.404	2	0.0671
RAPS (11+ years) Involvement Primary Caregiver	1821	275.030	6	0.000	**0.167**	**2**	**0.9197**	**0.884**	**2**	**0.6427**
RAPS (11+ years) Involvement Secondary Caregiver	1133	134.822	6	0.000	**1.146**	**2**	**0.5639**	**0.074**	**2**	**0.9637**

	N		df			df			df	
RAPS (11+ years) Autonomy Support Primary Caregiver	1817	**0.362**	**1**	**0.5472**	**1.674**	**2**	**0.4329**	**0.817**	**2**	**0.6647**
RAPS (11+ years) Autonomy Support Secondary Caregiver	1130	**0.030**	**1**	**0.8628**	**1.292**	**2**	**0.5242**	**1.317**	**2**	**0.5176**
RAPS (11+ years) Structure Primary Caregiver	1817	102.924	3	0.000	7.048	2	0.0295	**0.717**	**2**	**0.6988**
RAPS (11+ years) Structure Secondary Caregiver	1131	64.253	3	0.000	6.559	2	0.0377	**0.593**	**2**	**0.7436**
SF-12 Mental Health of Caregiver	5491	103.969	1	0.000	8.019	2	0.0181	7.788	2	0.0204
SF-12 Physical Health of Caregiver	5491	45.927	1	0.000	**4.008**	**2**	**0.1348**	5.333	2	0.0695

* Note: Constraints that could not be rejected are highlighted in bold: 2 uncorrelated errors, 7 constant error variance, and 10 constant true score variance