


2012

## **Internal Consistency of the Self-Perception Profile for Children: Using Covariance Structure Modeling to Overcome the Limitations of Cronbach's $\alpha$**

Ian Cero  
*Minnesota State University, Mankato*

Follow this and additional works at: <https://cornerstone.lib.mnsu.edu/etds>

 Part of the [Clinical Psychology Commons](#), [Quantitative Psychology Commons](#), and the [Statistics and Probability Commons](#)

---

### **Recommended Citation**

Cero, I. (2012). Internal consistency of the self-perception profile for children: Using covariance structure modeling to overcome the limitations of Cronbach's  $\alpha$ . [Master's thesis, Minnesota State University, Mankato]. Cornerstone: A Collection of Scholarly and Creative Works for Minnesota State University, Mankato. <https://cornerstone.lib.mnsu.edu/etds/23>

This Thesis is brought to you for free and open access by the Graduate Theses, Dissertations, and Other Capstone Projects at Cornerstone: A Collection of Scholarly and Creative Works for Minnesota State University, Mankato. It has been accepted for inclusion in All Graduate Theses, Dissertations, and Other Capstone Projects by an authorized administrator of Cornerstone: A Collection of Scholarly and Creative Works for Minnesota State University, Mankato.

Internal Consistency of the Self-Perception Profile for Children:  
Using covariance structure modeling to overcome the limitations of Cronbach's  $\alpha$

By  
Ian Cero

A Thesis Submitted in Partial Fulfillment of the  
Requirements for the Degree of  
Master of Arts  
In  
Clinical Psychology

Minnesota State University, Mankato

Mankato, Minnesota

May, 2012

Internal Consistency of the Self-Perception Profile for Children: Using covariance structure modeling to overcome the limitations of Cronbach's  $\alpha$

Ian Cero

This thesis has been examined and approved by the following members of the thesis committee.

Dr. Sarah Sifers, Advisor

Dr. Lisa Perez

Dr. Cindra Kamphoff

### **Acknowledgements**

It is difficult for me to over-state my gratitude toward the individuals who made this thesis possible. I am especially thankful for the feedback from advisor, Dr. Sarah Sifers, whose guidance has been indispensable to the completion of this thesis, in addition to many other research projects and my professional development, more broadly. I also wish to thank Dr. Lisa Perez for offering her expertise psychometric theory and for introducing me to the covariance structure model, a tool that I expect to utilize for many years to come. I am further grateful for the critiques and suggestions provided by Dr. Cindra Kamphoff, which enabled the arguments presented in this project to resonate more clearly and effectively. Lastly, I am indebted to many other individuals who provided substantial encouragement and emotional support, without which this project may never have been completed. Thank you.

**Abstract**

Self-perception is linked to a variety of psychosocial outcomes and its measurement has become a priority across a several disciplines. The Self-Perception Profile for Children (SPP-C) is commonly utilized to measure both global self worth and several important sub-domains of self-perception. Although much research has suggested this instrument possesses good internal consistency, previous investigations have primarily employed Cronbach's  $\alpha$  to estimate the stability of responding across items. This represents an important limitation, as  $\alpha$  is vulnerable to mis-estimation in the presence of correlated errors and non- $\tau$ -equivalent indicators, neither of which have been ruled out for the SPP-C. The present investigation initially examined the SPP-C responses from 106 girls, aged 8-12 to assess whether the assumptions underlying Cronbach's  $\alpha$  could be justified for this instrument. The investigation then re-estimated the internal consistency of the SPP-C using a covariance structure modeling approach. Results show that not a single scale of the SPP-C met all of the requirements for accurate estimation with  $\alpha$ . In two cases,  $\alpha$  was found to be meaningfully different from the reliability estimated with more durable methods. Lastly, the reliability of the Physical Appearance (PA) sub-scale was not significantly greater than the .70 cutoff for use as a research instrument. Discussion is centered on the appropriateness of Cronbach's  $\alpha$  for estimating the reliability of the SPP-C and recommends revision of the PA sub-scale.

**Table of Contents**

List of Tables .....6

List of Figures .....7

Methods .....27

Results .....33

Discussion .....35

References .....44

Appendix A: Tables .....49

Appendix B: Figures .....51

**List of Tables**

Table 1: PA items with method bias potential

Table 2: Item pairs with plausible error covariances

Table 3: Nested models evaluating the assumptions of Cronbach's  $\alpha$  five primary SPP-C sub-scales

Table 4: Nested models evaluating assumptions of Cronbach's  $\alpha$  for GS Subscale

Table 5: Reliability estimates and discrepancies of the SPP-C sub-scales

**List of Figures**

Figure 1: Path diagram relating the five domain-specific SPP-C latent constructs to sub-scale items

Figure 2: Path diagram relating GS latent construct to sub-scale items

Figure 3: CFA Reliability and Cronbach's  $\alpha$  estimates



Self-concept has been debated for over a century and has undergone several revisions since its inception. Initially, James (1890) argued that self-concept was the relationship between an individual's perceived competence in a given dimension compared to the significance and individual attributes to that same dimension. Coopersmith (1967) later responded with a uni-dimensional, social definition in which individuals compare their own competence to the competence of others. Contemporary models, however, understand self-concept as a series of self-evaluations across several distinct domains (Harter, 1993).

Harter (1985) argued that self-concept is comprised of five individual dimensions: behavioral conduct (i.e. self-control, ability to follow rules; BC), scholastic competence (SC), social acceptance (SA), athletic competence (AC), and physical appearance (PA). The aggregate of these domain-specific self-concepts can then be used to represent overall self-concept. In this way, multiple dimensions are combined to make a higher-order representation of self-concept. That is, self-concept is both multidimensional, in that it is comprised of distinct areas of competence, and hierarchical, in that specific competencies can be combined and treated as a general self-concept.

To quantify children's self-esteem, Harter (1982) developed the Perceived Competence Scale for Children, which assessed cognitive, social, and athletic competence. This scale also included a Global Self-Worth (GS) dimension to assess the general self-perception of children, rather than the individual sub-domains measured by its counterparts. The 28-item measure was later adapted to include sub-scales for behavioral conduct and perceived physical appearance also. After additional revisions to the original sub-scales, the instrument was renamed the Self-Perception Profile for Children (SPP-C; Harter, 1985). The instrument employs a forced-choice format wherein

children first choose between a set of statements (e.g. “Some kids would rather play outdoors in their spare time” vs. “Other kids would rather watch T.V.”) and then rate the degree to which their chosen statement is reflective of them (i.e. “sort of true for me” vs. “really true for me”). The updated version of this instrument is now composed on 36 items, with 6 items for each of the five domain-specific sub-scales and an additional 6 item sub-scale designed to assess GS.

The SPP-C has since been used for research in developmental, social, and clinical contexts (Cairns, McWhirter, Duffy, & Barry, 1990; Cross & Madson, 1997; Sciberras, Efron & Iser, 2011). It has also been used to assess children's experience in organized athletic activities and to examine self-concept in a variety of pediatric populations (Eapen, et al., 1992; Eriksson, Nordqvist & Rassmussen, 2008; Kapp-Simon, et al., 1992; Ridger, Fazey & Fairclough, 2007).

While its use across a variety of research and applied settings provides strong evidence for the importance of the SPP-C, the evidence of its internal consistency is limited by potential methodological problems. That is, nearly all previous attempts to estimate the internal consistency have employed Cronbach's (1951)  $\alpha$ , which relies on a set of restrictive assumptions rarely met in applied research (Brown, 2006). Further, violation of these assumptions can have marked effects on the behavior of the coefficient, artificially increasing or decreasing its estimation (Raykov, 2001). To address these issues, this paper will first describe the theory and assumptions of Coefficient  $\alpha$ , then explain how the SPP-C is likely to be affected by violation of these assumptions. Violation of Coefficient  $\alpha$ 's assumptions will then be formally tested and reliability will be re-estimated using confirmatory factor analytic (CFA) methods.

## Reliability

Reliability is simultaneously one of the most important and tedious considerations of psychological measurement. Understood as the agreement between raters, the consistency of responses over time, or as the stability of responding across items from the same test pool, the concept of reliability has a direct impact on a wide array of basic and applied research settings in the social sciences (Kline, 2005). The observed relationship between a set of variables can only be as strong as the measurement of the individual variables, themselves. Weakly measured constructs, mathematically, must have weakly measured relationships.<sup>1</sup> For this reason, reliability forms the upper-bound for validity and its examination is a necessary prerequisite to drawing inferences from all types of measurement, but especially self-report data (Raykov, 2001).

Although it comes in many forms, reliability is most often analyzed from the perspective of Classical Test Theory (CTT; see Kline, 2005). Within this framework, measurements (e.g., data gathered from a self-report survey) are assumed to be contaminated with some amount error due to random and uncontrollable factors (e.g., testing environment, hunger, amount of sleep prior to measurement). Thus, an individual's score ( $Y_i$ ) on any given instrument can be understood as the composite of their true score on the construct of interest ( $T_i$ ) and the error in measurement ( $E_i$ ) observed in that particular administration of the survey. This is expressed:

$$Y_i = T_i + E_i \quad (1)$$

---

<sup>1</sup> Kline (2005) describes some exceptions (e.g., change scores) where the relationship between two variables can be higher than their respective reliability coefficients. However, in such situations, there is typically another methodological flaw affecting the measurement under consideration (i.e., limited variability)

When this model is extended to multiple scores, the observed score variances follow the same pattern:

$$\text{Var}(Y_i) = \text{Var}(T_i) + \text{Var}(E_i) \quad (2)$$

where the variance in observed scores ( $\text{Var}(Y_i)$ ) is the sum of variance in true scores ( $\text{Var}(T_i)$ ) and variance in measurement error ( $\text{Var}(E_i)$ ). The stability of responses on an instrument (i.e., reliability), then, is defined as the ratio between the variance associated with what we desired to measure and the overall variance that was observed. Put in CTT terms, reliability ( $r_{xx}$ ) is the proportion of true score variance relative to observed score variance; or equivalently, reliability is the portion of observed variance in scores that is *not* due to error. Thus,

$$\rho = \frac{\text{Var}(T_i)}{\text{Var}(Y_i)} = 1 - \frac{\text{Var}(E_i)}{\text{Var}(Y_i)} \quad (3)$$

As suggested above, reliability is a broad term that describes a wide variety of measurement properties that a scale can possess. Internal consistency is defined as the stability of responding across the items of a scale administered at one time point and, likely due to the fact that relatively little data are required to evaluate it, is the most commonly reported subtype of reliability (Kline, 2005). Given how ubiquitous internal consistency is, this form of reliability will be the focus of the remainder of this discussion. Hence, future use of “reliability” should be read “internal consistency.”

**Cronbach's  $\alpha$** 

Because the relative amount of true and error variance that comprise a total score cannot be directly examined, a variety of frameworks have been proposed to assess the internal consistency of psychological instruments. In the last several decades, however, none has received wider acceptance than Cronbach's  $\alpha$ , which is equal to the average of all possible split-half correlations:

$$\alpha = \frac{K\bar{c}}{\bar{v} + (K-1)\bar{c}} \quad (4)$$

where  $K$  is the number of items on an instrument,  $\bar{c}$  is the average covariance between items, and  $\bar{v}$  is average variance. (Cronbach, 1951). Though versions of  $\alpha$  were available prior to its famous exposition--Kuder and Richardson (1937) had created a version for dichotomous items and Guttman (1945) listed the average split-half correlation as part of a broader series of reliability estimates--history has attached Cronbach's name to it.

The popularity  $\alpha$  has accrued in recent years is likely owed to a variety of factors, not least of which is its availability in standard statistics software packages (e.g., SPSS; Borsboom, 2006). Cronbach's  $\alpha$  also avoids many of the pitfalls contained in earlier indices of internal consistency, like the split-half's potential to provide different results depending on how the halves are split. Rather, Cronbach pointed researchers to a simple, standard, and unique estimate of a scale's reliability. Despite its popularity though,  $\alpha$  is met with important limitations, including its sensitivity to the overall size of a scale, and vulnerability to mis-estimation in the presence of correlated errors or items that are not  $\tau$ -equivalent.

### **Effect of Scale Size**

First, Cronbach's  $\alpha$  is sensitive to factors that comprise reliability other than the true, error, and observed variance. As can be seen with only a brief glance at its formula,  $\alpha$  increases proportionate to the number of items included in a scale. This feature of the coefficient means that even poorly designed scales, if they are sufficiently large, could garner  $\alpha$  values equal to smaller scales with superior psychometric properties. While the CTT framework predicts that instruments with more items will generally be more stable due to the fact that they sample a construct more comprehensively, that is an argument to *expect* good reliability from longer scales; however, it is not an argument for why a reliability index should *automatically* increase with the number of items on a measure. Rather, reliability is defined in (3) as the ratio of true score variance to observed score variance and, as such, should vary only according to those two values. Thus, a primary disadvantage of Cronbach's formula is that its outcome is influenced by factors that are unrelated to the CTT formulation of reliability.

### **Effect of Correlated Errors**

In addition to problems explicit in the formula, more subtle problems have been observed in the measurement model that underlies Cronbach's  $\alpha$ . Specifically,  $\alpha$  relies on a set of relatively impractical assumptions that are often overlooked by researchers. First,  $\alpha$  is likely to misestimate reliability in cases when item-error variances are correlated, even when population-level data are available (Brown, 2006). This is because the formula for  $\alpha$  only evaluates the total score covariance between items, rather than the true-score covariance, which is the researcher's primary interest. Thus, if the error component of

scores on item  $i$  correlate with the error component of scores on item  $j$ , then the total covariance scores will be contaminated with measurement error and so will the resulting  $\alpha$ . Put another way,  $\alpha$  is an inaccurate estimate reliability if error covariances are meaningfully different from zero.

The mathematical demonstration of this problem is straight forward. If total item variances are comprised of true score and error variances as indicated in (2), then the total covariance between item scores ( $Cov(Y_{ij})$ ) must also be comprised of covariances between their respective true score covariance ( $Cov(T_{ij})$ ) and error covariance ( $Cov(E_{ij})$ ). Thus,

$$Cov(Y_{ij}) = Cov(T_{ij}) + Cov(E_{ij}) \quad (5)$$

When the right term of (5) is substituted into the formula for  $\alpha$ , equation (4), the effect of error covariances is illustrated:<sup>2</sup>

$$\alpha = \frac{K\bar{c}}{\bar{v} + (K-1)\bar{c}} = \frac{\sum_{i \neq j}^K [Cov(T_{ij}) + Cov(E_{ij})]}{v + (K-1) \sum_{i \neq j}^K [Cov(T_{ij}) + Cov(E_{ij})]} \quad (6)$$

As shown by formula (6), that value of Cronbach's  $\alpha$  depends, in part, on the relationship between error terms, and thus, can be inflated or deflated depending on the direction and magnitude of their correlation (Brown, 2006). Typically, as error covariance

---

2 Note that when  $i = j$ , the resulting value would be the covariance between an item and itself. Namely, it would be a variance, which has been accounted for elsewhere in the derivation of this formula. Thus, the  $i \neq j$  condition placed on  $\sum$  is simply meant to indicate that variances should not be included in the sum of covariances.

increases,  $\alpha$  will increase because observed score covariance increases along with error covariance. However, depending on the nature of the error-correlation,  $\alpha$  can also underestimate the true reliability of a scale. Thus, unless a researcher has been unusually careful and verified that the error covariance between items is negligible,  $\alpha$  will inaccurately estimate the true reliability of a scale, even at the population level (Raykov, 2001).

### **Effect of $\tau$ -equivalence Violations**

Perhaps even more common than correlated errors are violations of  $\tau$ -equivalence, which can also artificially inflate or deflate  $\alpha$  (Brown, 2006; Cortina, 1993; Lord & Novick, 1967; Raykov, 2001). To explain, the general goal of administering sub-scales is to locate a single value that approximates a participant's true score on some latent dimension (e.g., self-esteem, social acceptance, athletic competence). This score is typically acquired by taking the raw or weighted sum (average) of a participant's scores across the various items that comprise a sub-scale. Cronbach's  $\alpha$ , provides an estimation of how reasonable this practice is for a given sub-scale by evaluating the covariances between items. However, it does so by assuming that all items measure the underlying construct in the same way. This is called the assumption of  $\tau$ -equivalence, which requires that items of a scale all measure a latent construct in the same units, and that they differ only by an additive constant (i.e., intercept). Mathematically,  $\tau$ -equivalence states that an observed score ( $Y$ ) on item  $i$ , relative to item  $k$ , is equal to their true score, plus a constant ( $c_k$ ), plus measurement error.

$$Y_{ik} = (C_k + T_i) + E_{ij} \quad (7)$$



Put in practical terms, this assumption requires that for any multi-item scale, a one unit increase in the latent construct of interest must affect the same unstandardized (i.e., not z-score) unit-change on all items. For example, if one unit-increase in the latent construct is reflected by a two-unit increase in the first item of the scale, then a one unit-increase in the latent construct must also be reflected by a two-unit increase in all other items on the scale, plus a constant. The constant may seem counter-intuitive, but is merely meant to indicate that the items can differ in their means, as long as their units change to the same degree as the latent variable changes. Thus, two hypothetical items can have a mean of 2 and a mean of 3, but as long as both of them change at the same rate when scores on the latent variable change, they are considered  $\tau$ -equivalent.<sup>3</sup>

Scales that are not  $\tau$ -equivalent, called “congeneric” scales, possess items that can measure the latent construct of interest with the same degree of accuracy (i.e., their correlations with the latent variable are the same), but they differ by both an additive (i.e., intercept) and multiplicative (i.e., unstandardized loading or slope) constant. This means they would measure the latent variable on a different scale (e.g., in inches rather than centimeters), so a one-unit increase might mean a two-unit increase in the first item on a scale and a three-unit increase on the second. Mathematically, this is expressed:

---

3 Strictly, *truly*  $\tau$ -equivalent models require that indicators measure constructs on the same scale, with the same precision (Graham, 2006). This means that not only must the rate of change relative to the construct be the same among indicators (the “same scale” assumption), but the means are also not allowed to differ (the “same precision”). Mathematically, the additive constant that allows for means to differ among items does not affect the reliability of an instrument, and thus, is not usually considered. This has given rise to the “essentially”  $\tau$ -equivalent model, which, technically, is what is required for  $\alpha$  and is being discussed in this paper. However, because the distinction between “essential” and truly  $\tau$ -equivalent models will not affect Cronbach's  $\alpha$  and is not relevant for this analysis, I have chosen to drop the “essential” qualifier in the name of simplicity.

where a score on any item, relative to any other item, can be represented as a true score, plus a constant, plus error in measurement. However, in this case, the slope representing the relationship between true score and observed item scores differs depending on the item in question and a multiplicative constant,  $d_k$ , must be included in the model to account for this factor.

### **An Empirical Demonstration**

The effect of varying violations of  $\tau$ -equivalence on the performance of  $\alpha$  can be directly examined with a formula provided by Raykov (2001). However, this theoretically derived formula and the accompanying explanation are laborious and more detailed than is required for the current discussion. Further, the formula is theoretical and includes terms that cannot be directly observed, making it irrelevant for actual research programs. The more mathematically rigorous presentation is merely mentioned here to indicate that the impact of  $\tau$ -equivalence violations on Cronbach's  $\alpha$  can be proven theoretically and that empirically detected outcomes of those same violations are more than just artifacts of the observed data.

Although the relationship between  $\alpha$  and  $\tau$ -equivalence is mathematically complex, an intuitive explanation is easily attained with an empirical example. To demonstrate the effect of  $\tau$ -equivalence violations on the performance of  $\alpha$ , a set of hypothetical scales were created, similar to those described in Graham (2006). First, true scores ranging from 1-10 were randomly generated for a group of hypothetical participants ( $n = 100$ ). Then, a scale was constructed by randomly generating four sets of completely uncorrelated error terms for each participant ( $r_{ij} = 0$  for all items, where  $i \neq j$ ),

which the true score was added to. This resulted in five columns of data for each participant: one for their true score, which will be ignored for the remainder of the demonstration, and four columns representing their scores on different items (“ $X_{1-4}$ ”) of a hypothetical scale designed to assess their true score, but contaminated with some error.

The construction of these items meets all the assumptions outlined by CTT and resulted in a scale with a reliability of .92, which can be calculated directly with (3) because true and observed variances are both known in this case. To simulate the impact of  $\tau$ -equivalence violations, the true score of item  $X_4$  was multiplied by 7. This resulted in a nearly identical correlation with the true score ( $r_{\text{change}} < .01$ ), but a marked effect on the reliability estimate:  $\alpha$  dropped to .65.

To examine the effect of correlated errors, the  $\tau$ -violating item was first replaced with a normal one. However, when moderately correlated errors between items  $X_1$  and  $X_2$  were introduced to the simulated scale ( $r_{\text{error}} = .35$ ),  $\alpha$  increased again to .97. When the correlation between errors on those same items was increased ( $r_{\text{error}} = .60$ ), the  $\alpha$  dropped to .84. Moreover, if correlated errors and  $\tau$ -violations were both introduced to the scale simultaneously, the resulting  $\alpha$  was .50.

Thus, while  $\tau$ -violations typically drag  $\alpha$  down, correlated errors can either inflate or deflate an  $\alpha$  value. When multiple assumptions of  $\alpha$  are violated, what would otherwise be highly reliable scales yield  $\alpha$  coefficients well below the standard for acceptable measurement properties (i.e.,  $\alpha \geq .70$ ). These results are similar to other Monte Carlo analyses and show that a single item (meaningfully) violating the assumption of  $\tau$ -equivalence or a single (meaningful) non-zero error covariance can cause a scale with high reliability to yield an  $\alpha$  coefficient markedly different from its true psychometric properties (Graham, 2006; Raykov, 2001). Despite the fact that their importance is easy

to demonstrate, there is not good reason to believe that researchers check for  $\tau$ -equivalence and correlated errors, much less compensate for them (Brown, 2006).

This fact is especially striking given how restrictive these assumptions are. For example, similar wording on two items could easily produce correlated errors, owed to self-consistent responding (Podsakoff et al., 2003). Further, the units of measurement in psychological assessment (e.g., dichotomous response vs. Likert-type formats) are often arbitrary. The arbitrary nature of most psychological scale units indicates that, unlike the Central Limit Theorem, which suggests that sample data are fairly likely to come from normally distributed sample distributions (even when researchers do not verify this fact), there is no a priori reason to believe that previously constructed scales have met the primary assumptions of Cronbach's  $\alpha$ . That is, unlike in most cases of inferential statistics, theory does not provide reasons why previously generated results are likely to hold. In the case of reliability, researchers must verify assumptions directly.

Returning to the SPP-C, it is important to note that there are unique reasons for caution when interpreting previous investigations using  $\alpha$  as an estimate of internal consistency. Criticisms involving response-format confusion and method biases, along with previous factor analytic investigations all suggest that correlated errors are a potential concern for this instrument. Further, while potential invariance of the measurement model has been examined across many demographic variables (e.g., age, gender, ethnicity), the measurement model required for accurate Cronbach's  $\alpha$  estimates--the " $\tau$ -equivalent" model (see Brown, 2006)--has never been examined. As the above discussion shows, these issues provide reason to be cautious of previous reliability estimates of the SPP-C that were examined with Coefficient  $\alpha$ , which is nearly all of them. The remainder of this section will focus on factors that increase the SPP-C's risk

for correlated errors and then consider previous investigations into the internal consistency of the instrument, proper.

### **Critiques of the SPP-C**

Previously developed critiques of the SPP-C generally focus on the response-format, consistency of wording across items, and broader concerns about the factor structure of the instrument, including the presence of correlated errors. For example, Eiser, Eiser, and Havermans (1995) found that a noticeable proportion of children were confused by the two-step-forced-choice response format. This finding was later replicated when Item Response Theory (IRT) analysis found that 5.7% of respondents who completed the SPP-C had atypical response patterns (Meijer, Egberink, Emons, & Sutsma, 2008). Subsequent investigation, which included consulting teachers and re-administration of the scale to the atypically responsive participants, revealed that these response patterns were often associated with confusion about how to fill out the measure. These results suggest that, although the somewhat novel response-format is presumed to reduce social desirability bias, the additional cognitive demands created by the response format may outweigh reduction of socially desirable responses. This is especially concerning, given that no empirical support has been presented for a reduction in socially desirable responding due to this scale structure.

In addition to critiques related to the format of the SPP-C, there are also reasons to suspect this instrument may be at unique risk for correlated errors. Recently, Podsakoff et al. (2003) compiled a list of common sources of method bias, which are likely to give rise to correlated errors. While these sources are only *potential* sources of covarying errors and, as such, do not guarantee correlated error-terms are present, it is concerning

that many pairs of SPP-C items fall into at least one risk category and some fall into multiple categories. As shown in Table 1, the second, third and fourth items from the PA sub-scale are worded ambiguously, similarly, encourage self-consistent responding through their nested content. More directly, the features of the “different” body are not clear, all of these items include the word “different” in one of the forced-choices, and answers on one item are logical subsets of answers on another. To this last point, consider that each item is nested within another on the list. If a child likes her appearance (Item 22), she is also inclined to say that she likes her body as it is, along with her height and weight as they are. The content on each of these items is nested within the content of the one that follows it. This kind of nested structure puts children at risk for self-consistent responding because answering “yes” on one item implies a “yes” on another.

The reader may be inclined to defend these items along two lines. First, perhaps they are not *that* ambiguous, participants are likely aware that when they think about a “different” body that they should really be considering a superior one. Second, the fact that participants try to be logically consistent when they respond to these items might be taken as evidence the items measure the same construct. Thus, a child answers similarly on all three of these questions because the items all assess the same underlying construct. Consistent responding, then, might be a good thing.

To the first argument, it should be noted that many of the participants who are assessed with this instrument are young children and research has already demonstrated nearly 6% of them are confused by the instrument in some way or another. While some children might make that inference, it is not safe to assume all of them do. In fact, that is the point. As Podsakoff et al. (2003) explain, when items become ambiguous, participants resort their own interpretations, rather than the interpretation intended by the researcher.

That is, some children may interpret “different” as “better” and some may interpret it as “any different body, better or worse.” These different interpretations introduce new, and likely correlated, sources of error into the instrument, which is already confusing to a significant portion of its respondents

Regarding self-consistent responding, the argument here is *not* that covarying items are bad. On the contrary, strong correlations between items are quite important for accurate measurement of a construct. However, when items are covary for reasons unrelated to the construct of interest, correlated error terms emerge. As will be discussed below, this can have important implications for the reliability of a sub-scale and is worth avoiding.<sup>4</sup>

While the above analysis about wording on the SPP-C can be made by mere inspection of its items, it is worth noting that more rigorous investigation of this instrument's properties has yielded similar conclusions. Eiser, Eiser, and Havermans (1995) speculated that the factor structure of the SPP-C may be, at least in part, an artifact of similar wording on the BC, PA, and GS sub-scales. More recently, Egberink and Meijer (2011) employed both parametric and non-parametric IRT methods to analyze the behavior of each sub-scale. They found that the psychometric properties of PA and GS items actually improved when they were combined into one large scale. While it is counter-intuitive to view an increase in the measurement properties negatively, in this case it suggests potential problems with the sub-scales. Assuming each sub-scale measures a distinct domain of self-perception, then combining scales should reduce their overall functioning. The authors explain that, in this case, finding an improved scale after

---

4 It is important to note that the validity of inferences drawn from a scale is also affected by correlated error terms. However, discussion of such effects is beyond the scope of this manuscript (for an overview, see Podsakoff et al, 2003.).

two distinct sub-scales are combined is suggestive of repetitive wording across those sub-scales.

Other studies have investigated the SPP-C construct more broadly with factor analysis. Results of many investigations have shown general support for the original five-factor structure that Harter (1985) hypothesized at the advent of the measure (Boivin, Vitario, & Gagnon, 1992; Granleese & Joseph, 1993; Miller, 2000; Schumann et al. 1999; Van den Bergh & Ranst, 1998; Van den Bergh & Marcoen, 1999). There are, however, important caveats to the general pattern of results in these studies.

Notably, the factor structure has not been invariant across all demographic groups. Stewart, Roberts, and Kim (2010) performed an exploratory factor analysis of the SPP-C after it had been administered to a group of African American girls. Results showed that none of the originally proposed factors were replicated. In fact, one of their observed factors included items from four separate SPP-C sub-scales. Examination of the factor structure in this population suggests that, for African American girls, SC and BC are not differentiated. It is notable that principle components analysis has yielded similar results. Schumann et al (200) demonstrated that although reliability and component structure improved over time, SC and BC remained undifferentiated for African American girls when they reached age 12. These results did not change when high- and low-SES groups were analyzed separately. It is worth noting that neither of these studies are “true” tests of non-invariant measurement, like confirmatory factor analysis (CFA), but the results do suggest that non-invariant measurement across ethnicities is a risk that should be investigated.

In addition to factor invariance across ethnic groups, research has suggested that the factor structure of the SPP-C varies as a function of gender. Van der Bergh & Ranst



(1998) performed separate confirmatory factor analyses for boys and girls. Although they found that their final models both included five factors with the same items loading on each factor, they were only able to achieve those results by relaxing different sets of assumptions. For boys, a five-factor model achieved adequate fit when the SC, SA, and BC error terms were allowed to correlate. However, for girls, adequate fit for a five-factor model was achieved only after relaxing error covariance restrictions SC, SA, PA, and BC. Taken together, these results suggest that there are likely to be meaningful error covariances between items on the SPP-C. Further, the pattern of these correlated errors is likely to differ from group to group.

### **Research on the Reliability of the SPP-C**

Although some of the arguments presented above highlight concerns about the psychometric properties of the SPP-C, it is important to note that the reliability data is generally favorable. Harter's (1982; 1985) initial validations of the instrument indicated that its sub-scales were generally stable, with  $\alpha$  values ranging from .71 and .86 across multiple age groups. Since then, many studies, typically focused on the factor-structure and validity of the SPP-C, have generally found internal consistency results similar to Harter's original investigation (Muris, Meesters, & Fijen, 2003; Schumann, 2000; Van Dongen, Koot, & Verhulst, 1993).

However, the reliability varies markedly across sub-scales, with the BC sub-scale typically fairing the worst and PA typically fairing the best (Worth, Gavin, & Herry, 1996). Second, though Harter's original results indicated that all scales yielded satisfactory  $\alpha$  values ( $\alpha \geq .70$ ), some studies have observed reliability estimates below the standard threshold (Hess & Peterson, 1996). Third, many sub-scale  $\alpha$  coefficients have

also been shown to vary across demographic variables, including age, gender, and ethnicity (Boivin, Vitario, & Gagnon, 1992; Eapen, Naqvi, & Al-Dhaheeri, 2003; Stewart, Roberts, & Kim, 2010; Van Der Bergh & Van Ranst, 1998). Fourth, Cronbach's  $\alpha$  estimates have also been shown to vary across administrations (Shevlin, Adamson, & Collins, 2003).

Though nearly all studies examining the reliability of the SPP-C have relied on coefficient  $\alpha$ , there are three investigations that have employed more advanced methods. Shevlin, Adamson, and Collins (2003) tested the higher-order measurement invariance of the overall scale through four administrations. Results showed that the relationship between each sub-scale and the higher-order self-perception factor proposed by Harter (1985) was not invariant across time. Worth Gavin and Herry, (1996), in addition to estimating reliability with Cronbach's  $\alpha$ , also considered the squared multiple correlations between each item and its underlying factor, as estimated by a confirmatory factor analysis (CFA). Wide variability was observed in the correlation between indicators and their factors, with only three items crossing acceptable thresholds ( $R^2 \leq .50$ ) across all age-groups in their sample (Byrne, 1989). Further, some items failed to cross this threshold in any age-group, this was most commonly observed in items on the BC subscale. Lastly, Van Der Bergh and Van Ranst (1998) examined changes in reliability across age and gender, as part of a larger analysis of the factor-structure of the SPP-C. Model fit significantly improved when error variances were not restricted to be the same across age and gender, which was interpreted as a change in reliability over time. While this interpretation makes some sense, increasing error variances suggest less of the observed item variance is comprised of true variance. This is not a formal test of reliability change. These results do not confirm unstable reliability; they only suggest that

it is likely.

Finally, although it is difficult to find evidence of  $\tau$ -equivalence violations without formally testing them, examining the confidence intervals for factor loadings provided from previous CFA investigations suggests that the units of measurement may be different between items (Boivin, Vitario, & Gagnon, 1992). Additionally, while research on the SPP-C has never constrained unstandardized loadings to equality, the formal test of  $\tau$ -equivalence, Shevlin, Adamson, and Collins (2003) did constrain some loadings to be equal over time. The authors found that fit significantly improved as equality constraints were lifted, indicating that the measurement properties of the scales changed over time. Though these results do not directly indicate that  $\tau$ -equivalence was violated, the fact that item relationships are unstable over time provides a reason to evaluate whether they are even initially stable at a single time point. That is, if a measure is observed to have low test-retest reliability, there could be many reasons for this, but the most obvious candidate is poor internal consistency. Likewise, if factor loadings are non-invariant across time points, they may be invariant within a single time point as well.

### **The Present Study**

Though the reliability of the SPP-C has generally been replicated, there is reason to be cautious about the estimates produced by previous research. As mentioned above, most studies estimated internal consistency with Coefficient  $\alpha$ . However, there is evidence that the assumptions required for this type of estimation may not have been met. Thus, the present study consisted of three phases. In the first phase, potential error covariances were located. Next, the SPP-C's ability to meet the assumptions of Cronbach's  $\alpha$  was formally tested with nested hierarchical confirmatory factor models.

Finally, reliability of the SPP-C sub-scales was estimated with both Cronbach's  $\alpha$  and CFA methods and compared.

## **Methods**

### **Participants**

Participants were a group of preadolescent girls ( $N = 106$ ) from central Minnesota, ages 8 through 12, who were enrolled in an after school program, Girls on the Run (GOTR). This preventative program runs for 12-weeks, with two sessions a week, and uses an experience-based curriculum to help girls maintain physical, emotional, and spiritual health. Preliminary research suggests the program is likely to improve body-satisfaction, eating attitudes and behaviors, and self-esteem (DeBate & Thompson, 2005). Participants were predominantly Caucasian. Parental consent and participant assent were obtained for all individuals in this study.

### **Design**

Data was collected from participants in eight separate iterations of the GOTR program, which spanned four years (two iterations per year). In each wave of data collection, research assistants administered self-report measures individually to participants. Responses were examined immediately after forms were filled out, and participants were asked to clarify ambiguous answers. Research assistants were available to answer any questions elicited by the participants over the course of survey administration. In cases where participants' reading skills were limited, survey items were read to them. While both pre-test and post test data were gathered, change in reliability over time is beyond the scope of this discussion. Thus, only pre-test data were evaluated in this investigation.

## Measures

As described above, the SPP-C (Harter, 1985) assesses a child's self-perception across five different sub-domains. Each sub-scale contains six questions and samples a different domain of self-perception, with the exception of the GS scale, which examines participants' general self-perception. Following arguments laid out in Shevlin, Adamson, and Collins (2003), the GS sub-scale was not considered in this investigation because it is not part of the general factor structure of the SPP-C proposed by Harter (1985). The SPP-C has been used in a wide variety of populations and has shown satisfactory factorial and convergent validity (Van Dongen et al., 1993).

## Analysis

**Phase 1.** To address this problem, the present investigation consisted of three phases. First, because there is no exploratory test to detect correlated errors in a measurement model, and because relying solely on modification indices from CFA software risks over-fitting the model of this instrument without substantive grounds (Brown, 2006), a pilot investigation was conducted to locate sets of items within each sub-scale that were likely to have correlated errors. To guard against over-fitting, this pilot investigation employed both an empirical and theoretical inclusion/exclusion criteria. That is, to be included in the next phase of the study, there must be empirical evidence that relaxing an error covariance would significantly improve fit, but also that the error covariance could be justified on the basis of previous analysis regarding correlated errors.

Initially, a CFA was performed for each sub-scale of the SPP-C separately and

modification indices were requested. Only error covariances with a modification index close to, the critical value for a  $\chi$ -difference test with 1 degree of freedom (3.84), were selected for further analysis. Separate estimation was indicated in this case because modification indices suggesting error covariances across sub-scales would not meaningfully impact the results of this investigation. This is because reliability is typically estimated for each sub-scale separately, so a cross-scale error covariance would not affect the calculation of reliability of an individual sub-scale. More directly, the goal of this procedure was to locate the plausible error covariances *within* a sub-scale and fitting all sub-scales simultaneously in one model would point to potential between scale covariances, as well as within scale covariances. Thus, models were fit separately.

After potential error correlations were detected, the item pairs involved in each potential error covariance were further inspected to see if it fell into any of the common categories of method biases and correlated errors described by Podsakoff et al. (2003). Only errors which had significant modification indices and fell into a common category of method bias were included in the next phase of the study.

It may be remarked that this process is backwards, that the theory portion of an investigation should precede the empirical testing. First, this pilot investigation is not a formal test of correlated errors. Rather, it was designed to suggest candidates that could later be tested with more formal methods. Along these lines, and consistent with explanations given by Brown (2006), modification indices were interpreted as evidence of a *potentially* meaningful error covariances, not as formal significance tests. Second, the number of possible item pairs grows factorially with the size of an instrument. In the case of the 36-item SPP-C, for example, there are 630 novel item pairs. Even when only within-scale error covariances are considered, 90 unique pairs would still need to be

investigated individually. Analyzing modification indices first is several times more efficient. Further, given that the candidate error covariances will be re-tested using confirmatory methods in phase two, the probability of a false positive derived from the pilot procedure being included in the final reliability calculation is low.

It may also be remarked that the empirical cutoff for modification indices ( $\approx 3.84$ ) does not provide a clear standard for inclusion and exclusion. This is intentional and consistent with the appropriate interpretation of modification indices, which are only best guesses at what actual change in fit *might* be, if a constraint were relaxed. Given what a modification index actually points to, providing a hard-cutoff, in the style of hypothesis testing, is not justified. Again, item pairs selected at this portion of the study were also subjected to two more analyses – a theoretical consideration and hierarchical confirmatory factor analysis -- before they were included in reliability estimation.

**Phase 2.** Following the procedure outlined in Brown (2006), a series of hierarchically nested CFA models were used to test for violations of  $\tau$ -equivalence and correlated errors. These nested models allow a researcher to compare the relative fit of adjacent models in the hierarchy with  $\chi^2$ -difference tests. Testing began with the least restrictive model and sequentially constrained each sub-scale to  $\tau$ -equivalence, followed by sequentially constraining, the error covariances gathered from Phase 1 to zero. When a significant difference in fit was observed, the constraint was discarded. If a new model was tested immediately after another model had been shown to be inappropriate, the new model was compared to the most recent model with appropriate fit. The models and their descriptions are available in Table 3. Hierarchical models were tested in R using the OpenMx package (Boker et al. 2011).

**Phase 3.** The final phase involved both reliability estimation with Cronbach's  $\alpha$ ,

performed in PASW Statistics 18, and more robust CFA procedure performed in LISREL 8.8 (Jöreskog & Sorbom, 2006; Raykov, 2001; 2002). This framework calls for three phantom (i.e., “dummy-coded”) latent variables, which are constrained to be equal to (a) the total variance of a construct, (b) the squared sum of unstandardized factor loadings, and (c) the resulting reliability created by dividing the squared sum of factor loadings by the total variance. Adding these phantom variables to the (already estimated) measurement model will have no impact on the overall fit of the final model because they are specified to have no impact on the covariance structure of the data. This procedure initially provides point-estimates of reliability, but Raykov (2002) has provided methods for estimating standard errors, and thus, confidence intervals for the reliability of an instrument. According to Raykov, the standard error of these reliability estimates is equal to:

$$SE_{\rho} = \sqrt{D_u^2 Var(u) + D_v^2 Var(v) + 2(D_u)(D_v)Cov(u, v)} \quad (9)$$

where  $u$  equals the sum of unstandardized factor loadings for a sub-scale,  $v$  is the sum of error variances on that same scale, and  $D_u$  and  $D_v$ , are the partial derivatives of the scale reliability estimate ( $r_{xx}$ ), with respect to  $u$  and  $v$ . It should be noted that the partial derivatives for  $u$  and  $v$  can be also be calculated with the following formulas:

$$D_u = \frac{(2uv)}{(u^2+v)^2} \quad (10)$$

$$D_v = \frac{u^2}{(u^2+v)^2} \quad (11)$$

Once  $SE(r_{xx})$  for each sub-scale was calculated, confidence intervals were



constructed. These intervals were inspected for two features. First, intervals were examined to see if they contained the .70 minimum cutoff for appropriate a research instrument (Nunnally, 1978), if the interval contained this value, it was concluded the reliability of the instrument was not significantly greater than the minimum cutoff. Second, each scale's reliability interval was inspected to see if it also contained the previously calculated Cronbach's  $\alpha$  estimate belonging to that same scale.

It should be noted that Cronbach's  $\alpha$  has its own distribution and asking whether  $\alpha$  is contained within the new reliability interval is *not a hypothesis test of whether these estimates are significantly different*. That would be analogous to creating confidence intervals to compare slopes from an ordinal regression to an ordinary least-squares (OLS) regression. The fact that the confidence intervals are non-overlapping is not a test that they are “significantly different” because they come from different sampling distributions. However, if the assumptions of OLS regression are violated, but the ordinal regression assumptions are not, then the fact that the OLS regression slope's confidence interval does not overlap the interval the ordinal regression slope may be interpreted as evidence – though not confirmation – that the OLS regression should be interpreted with caution. The same is the case here, that fact that  $\alpha$  estimates are not contained within the reliability interval is *not* confirmation that they are significantly different from the true reliability. However, it does provide evidence that we should be skeptical of  $\alpha$ , when it is not contained within the newly generated reliability interval.

A final note on the following analysis, the GS sub-scale is not considered part of the general factor structure of the SPP-C (Harter, 1985). Because of this, all procedures described in Phase 2 and Phase 3 were performed separately for this sub-scale. That is, it was tested for  $\tau$ -equivalence and correlated errors in a separate hierarchy of nested

models that did not include any of the other scales. Likewise, when its reliability was estimated, no indicators or latent variables from other scales were included in the model.

## Results

**Phase 1.** Examination of the modification indices from individually fit sub-scales suggested relaxing assumptions for 19 total error covariances. However, only 11 of these appeared to fall into at least one common category of method bias. These remaining 8 item-pairs, shown in Table 2, were considered to be plausible candidates for within-sub-scale error covariances, and were formally tested in the next phase of the primary study.

**Phase 2.** As shown by Table 3, the least restrictive model of the five domain-specific SPP-C sub-scales demonstrated satisfactory fit,  $\chi^2(386) = 578.86, p < .001$ , RMSEA = 0.062, TLI = 0.929, CFI = 0.937 and are near the ranges proposed by Hu and Bentler (1999). As expected, fit declined slightly as additional restrictions were imposed ( $\chi^2(407) = 604.77, p < .001$ , RMSEA = 0.064, TLI = 0.931, CFI = 0.935), though the final, most parsimonious model remained satisfactory. Significant decrements in fit, as assessed by  $\chi^2$ -difference tests, were observed when  $\tau$ -equivalence was constrained for AC, but not for any other sub-scale. Further, constraining error covariances to zero led to a significant decrement in fit for all potential cases derived from Phase 1, except the error covariance between Item 2 and Item 8 from the SA sub-scale were constrained. This error covariance was thus excluded from the reliability calculation in Phase 3.

As shown in Table 4, the pattern of results was similar when GS was estimated on its own. Both the least restrictive model ( $X^2(7) = 3.82, p < .799$ , RMSEA = 0.000, TLI = 1.000, CFI = 1.000) and the final model ( $X^2(13) = 8.17, p < .832$ , RMSEA = 0.000, TLI = 1.000, CFI = 1.000) resulted in satisfactory fit.  $\chi^2$ -difference tests did not reveal a

significant decline in fit when factor loadings were constrained to equivalence, nor did constraining the error covariance between Item 12 and Item 36 to zero. However, significant decrement in fit was observed when the error covariance between Item 24 and Item 30 was constrained to zero. Thus, the GS sub-scale is  $\tau$ -equivalent, but contains a correlated error.

Parameter estimates for the final models are depicted in Figure 1 and Figure 2. Note that in cases where error covariances were non-significant, they were still used in the calculation of reliability because they were shown to contribute to the overall fit of the model. However, given their small size, they are unlikely to affect the reliability estimation to a meaningful degree. In summary, all sub-scales are  $\tau$ -equivalent, except the AC sub-scale. Further, all sub-scales, except the AC sub-scale, contain at least one correlated error. Taken together, results suggest that every sub-scale of the SPP-C violates at least one assumption of Cronbach's  $\alpha$  to some degree.

**Phase 3.** Results of the two reliability analyses are shown in Table 5 and Figure 3. Consistent with much of the previous research on this instrument, all Coefficient  $\alpha$  estimates are satisfactory. Moreover,  $\alpha$  estimates are contained within the new reliability estimates in all but two cases. For the SC sub-scale, Coefficient  $\alpha$  is below the probable range of reliability. In this case,  $\alpha$  appears to be deflated because the error correlation between Item 1 and Item 19 is negative. Thus, the denominator of the CFA reliability formula is smaller than would otherwise be estimated with  $\alpha$ . It is worth noting that, despite the fact that  $\alpha$  falls outside the probably reliability range, this is only to a small degree  $\square$  is an *underestimate*.

The same cannot be said for the PA sub-scale, whose  $\alpha$  estimate falls far above the probable range of reliability as estimated by Raykov's (2001; 2002) procedure. Moreover,

the confidence interval for the reliability of the PA sub scale overlaps .70, suggesting it is not significantly better than the minimum cutoff for a research instrument (Nunnally, 1979). Taken together, these results suggest that the applied researcher should be cautious of estimates of the internal consistency of the PA sub-scale that used  $\alpha$ , but should also be cautious when interpreting results from the PA sub-scale more generally, as it does not clearly cross the minimum threshold for use as a research instrument.

### **Discussion**

Results suggest that each sub-scale of the SPP-C violates at least one of the assumptions required for Cronbach's  $\alpha$  to be an accurate estimator of reliability. Five scales – GS, SA, PA, BC, and SC – were observed to contain at least one non-zero error covariance. Further, although the AC sub-scale contained no correlated errors, this scale violated the assumption of  $\tau$ -equivalence. Thus, in this population, not a single sub-scale on the SPP-C met all the requirements for accurate estimation of reliability with Cronbach's  $\alpha$ .

While the results suggest that the assumptions underlying  $\alpha$  may be too restrictive for this scale, it is somewhat surprising that more violated assumptions were *not* detected. Namely, the fact that only one of the sub-scales of the SPP-C violated  $\tau$ -equivalence is impressive, given that there was not an a priori reason to believe that such a strict standard was likely to be met. As will be discussed in more detail below, it is further impressive that the one violation of  $\tau$ -equivalence that *was* observed was smaller than previous Monte Carlo analysis (e.g. Raykov, 1997) had even tested. This suggests that the fact that the AC sub-scale is not  $\tau$ -equivalent is unlikely to represent a meaningful practical challenge to the utility of the scale, even though this finding is statistically

significant.

When Cronbach's  $\alpha$  and CFA reliability estimates were compared in Phase 3, two  $\alpha$  estimates were found to be outside the plausible range of reliability for their respective sub-scales. In the case of the SC sub-scale,  $\alpha$  was below the reliability confidence interval. However, this was only to a small degree, and is unlikely to affect a researcher's confidence in the performance of that scale. In fact, as the results suggest  $\alpha$  was under-estimating the performance of this scale, confidence in the precision of its measurement should increase compared to previous research in this population, if only modestly.

For the PA sub-scale, however, Coefficient  $\alpha$  was markedly higher than the upper-bound of the reliability confidence interval. The reason for  $\alpha$ 's substantial over-estimation, in this case, is likely the three moderately sized error covariances that were detected on this scale. These results are not entirely unexpected, given previous research suggesting that the high agreement between items of the PA sub-scale may have been influenced by item characteristics unrelated to the construct of interest (e.g. similar wording; Egberink & Meijer, 2011).

It is important to consider, after locating violated assumptions for every sub-scale, why more outlying  $\alpha$  estimates were not detected. To explain, it is observed that the amount that  $\alpha$  over- or under-estimates reliability is a function of the number of offending items on a sub-scale, relative to the overall sub-scale length, and a function of the degree to which its assumptions have been violated (Raykov, 2001; Raykov, 1997). Given this, it is worth noting two things about the findings from Phase 2, before discussing the implications of the reliability estimates yielded in Phase 3.

First, in most cases a small number of items from each sub-scale involved violated assumptions of Coefficient  $\alpha$ . For example, on the SC sub-scale, only one error

covariance was detected. Additionally, although the AC sub-scale was found not to be  $\tau$ -equivalent, this appears to be influenced largely by the discrepancy between Item 3 and Item 33, whose loadings were respectively much lower and higher than their counterpart items. The remainder of the items on this sub-scale appear relatively similar to one another and are thus likely to have a negligible effect on  $\alpha$ . Given that, in most cases, a majority of items on each sub-scale were consistent with the assumptions underlying Cronbach's  $\alpha$ , it follows that these same scales should have only minor discrepancies between  $\alpha$  and the CFA-estimated reliability.

Second, in a majority of cases, the degree to which Coefficient  $\alpha$ 's assumptions were violated was minimal. For example, in the case of the non- $\tau$ -equivalent AC sub-scale, the largest and smallest loadings differ by a factor of 1.83, which is less than the smallest ratio of loadings that were evaluated in previous simulation studies. In fact, the smallest  $\tau$ -equivalence violation that was tested in the Monte Carlo analysis performed by Raykov (2001) was 2.00. Further, the results achieved by Monte Carlo analysis are similar to those observed in this study, when similar magnitudes are compared. As another example, correlated errors were detected on the BC sub-scale, but these correlations were fairly small and unlikely to influence  $\alpha$  to a great degree. When these factors are considered in aggregate, it is clear that Coefficient  $\alpha$  estimates would likely be different from the CFA estimates, but this difference should be small in most cases.

Even more important than the relative location Cronbach's  $\alpha$  to the CFA reliability, however, is the objective location of the new reliability estimate's confidence interval. Specifically, the confidence interval for the reliability of the PA sub-scale contained .70. This threshold, recommended by Nunnally (1979), is widely accepted as the minimum threshold for use as a research instrument (Lance, Butts, Michels, 2006). The fact that the

PA sub-scale reliability confidence interval is not distinct from this threshold, then, is a major concern. Given that there is not evidence that the PA sub-scale is significantly better than the minimum acceptable reliability in this sample, then, it suggested researchers should be highly cautious when interpreting previous results from this sub-scale in this population. Moreover, given the of lack of evidence this scale has crossed the minimum threshold, and given that research in another psychometric framework (i.e. IRT) has already raised suspicion about the behavior of items on this portion of the SPP-C, it is recommended the PA sub-scale be revised. Revision should be seen as a unique priority for this sub-scale of the SPP-C, which has the strongest correlation with global self-concept of all the sub-scales (Harter, 1985).

Fortunately, revision of this scale is likely to be straightforward. Because a convergence of evidence indicates that its deficits appear to be the result of similar wording across some of its items, simply removing some of the items of the PA sub-scale may improve its reliability. Thus, depending on the nature and results of the revision, the newly revised PA scale may have the potential to be simultaneously shorter and more reliable than the present version.

While the actual revision of the PA sub-scale is beyond the scope of the present analysis, it is noted here that removal of Item 10 and Item 16 would likely result in the greatest initial improvement in the utility of this scale. To explain, because the content of Items 10 and 16 is nested within the content of Question 22 (see introduction), removing the former two items would eliminate the correlated errors without major effect on the content of the scale. It is conceded that a comprehensive revision of the PA sub-scale will likely involve more than deletion of offending items. However, the initial removal of Items 10 and 16 would, at very least, reduce the risk that extraneous correlations between

items on the PA sub-scale are confound current findings.

It may be objected that is approach would result in a PA sub-scale that is shorter than all of its counterparts, throwing off the ‘balance’ of the scale. First, recall that the length of a scale is not directly related to its “true” reliability, only to its  $\alpha$ -value, which has been observed to mis-estimate the reliability PA sub-scale already. Thus, the scale’s length can be reduced, while reliability is preserved. Second, as long as all sub-scales yield reliable and accurate measurements of their underlying constructs, there is no further advantage accrued by requiring that all sub-scales be of equal length.

Turning to the remaining sub-scales, it is worth noting that although the .70 threshold is generally treated as the minimum acceptable reliability, it is sometimes misinterpreted. Nunnally's (1979) actual suggestion is that the .70 cutoff is acceptable only for “early stages of research,” but in “basic research...a reliability of .80 for the different measures is adequate” (p. 245-246). This position is seconded by Carmines and Zeller (1979) who claim that “As a general rule...reliabilities should not be below .80 for widely used scales” (p. 51). Given that the SPP-C is a mature scale that has already undergone substantial revision and is used widely across a variety of sub-disciplines, .80 is likely to be a more appropriate standard for evaluation. Additionally, if we adjust the standard for acceptable reliability for the sub-scales to this stricter threshold, only the BC sub-scale demonstrates acceptable reliability. None of the other sub-scales are not significantly above the more appropriate threshold.

Preemptively, it is conceded that no single cutoff is sufficient for the evaluation of all instruments (Lance, Butts, Michels, 2006). Depending on the intentions and precision required by the researcher, the reliability of the remaining SPP-C sub-scales may still be adequate. Thus, unlike the PA sub-scale, no further recommendations for revision are



made for the remaining scales.

It is reiterated, however, that a reliability .70 is the proposed minimum standard for *any* purpose (Nunnally, 1979). Thus, whereas the adequacy of many of the SPP-C scales can be interpreted flexibly and in the context of specific research goals, this is unlikely to be possible in the case of the PA scale. In its current state, the estimated range of reliability for the PA sub-scale is too low to be a matter of research context and would benefit from reworking.

### **Limitations**

As a caveat, it is mentioned that no one reliability cutoff is appropriate in all situations (Lance, Butts, & Michels, 2006). Depending on the researcher's intentions, the reliability of many of the SPP-C sub-scales may be sufficient for some research programs. However, the reader is reminded that there is special reason to be concerned about the reliability of PA sub-scale in this population, given its reliability is much lower than the other scales and is not above the generally agreed upon *minimum threshold*, even for scales in development. Again, this finding was anticipated by previous research which has raised concern about what is really being measured by the PA sub-scale (Egberink and Meijer, 2011). Because the PA sub-scale has been in use for some time, and because this is one of the first studies that can account for error correlations among its items, replication of this finding in additional will be important before any firm conclusions about its limitations can be made.

Further limitations concern the design of the study. First, the procedure for selecting potentially correlated error terms was designed to reduce the risk of over-fitting the measurement model (i.e., to reduce the rate of false positives). For this reason, and

because there were a great number of item pairs that were not even evaluated for potential method biases, it is possible that meaningful error correlations were missed in subsequent phases of analysis. However, over 20% of all possible within-scale error correlations were in the first phase of this study. Further, because these item pairs were already the best empirical candidates to improve the fit of the model, it is unlikely that major error correlations were missed. Nevertheless, future research should endeavor to utilize a procedure for locating correlated errors that minimizes both false positives and false negatives, in an effort to clarify the behavior of the sub-scales in this instrument.

It should also be observed that the sample for this project is smaller than in many other CFA studies. While this is certainly a reason that more research should be conducted using this method, there are a few important defenses of the results yielded here. First, the primary goal of this study was not to confirm the factor structure of the SPP-C, rather it was to account for the negative impact of that structure on the estimation of reliability. Second, the general model for this instrument has already been confirmed, several times, across several groups. The probability that the final models for the main SPP-C scales and the GS sub-scale were confirmed in with meaningful errors is low.

Further, the minimum appropriate sample size for CFA is not absolute and rules of thumb are known to have poor generalizability (Brown, 2006). They depend on factors unique to the sample and study design. In this case, relationships between indicators and their common factors were known to be relatively strong before hand, and the risk of misspecification is low.

With the exception of the often ignored  $\chi^2$  goodness of fit tests, all fit indices are within acceptable ranges. Despite the small sample, there is both statistical evidence and evidence from previous research that the models are correctly specified. Shevlin,

Adamson, and Katrina (2003) also performed a longitudinal evaluation of this instrument using a CFA model that was four times as complex as the one performed here, yet only 50 additional participants were used. All fit indices observed in that were still satisfactory.

Lastly, in an attempt to be sensitive to the risk of mis-estimation that comes with sample size, I have reported confidence intervals that contain ranges of likely values for reliability. Further, standard errors have also been presented for all parameter estimates in the final models, to give a sense of the potential dispersion of parameter values in the population. These strategies and the arguments presented above do not remove the implications of small samples for CFA and future research should endeavor to recruit larger groups of participants in an analysis such as this. However, confidence intervals and standard errors should mitigate some of the risk that the reliability calculated from the sample is not exactly the same as that of the population.

A final note, the sample utilized in this study was entirely female and predominantly Caucasian, which would normally limit the generalizability of the findings. However, as discussed above, the factor structure of this instrument changes slightly across different demographics. For this reason, a homogenous sample is actually a strength of the study because it reduces the risk that measurement bias has contaminated the results. Though future research should attempt reproduce this findings, it is advised that replication of the present findings ought to be conducted keeping the implications that demographics have on the behavior of the SPP-C in mind. That is, because it is possible that the patterns of error correlations and  $\tau$ -equivalence among items of the SPP-C sub-scales vary across ethnicity, and gender, attempts at replication of this study are encouraged to draw participants from a population with similar demographic characteristics to those used in this study. Likewise, researchers utilizing

the SPP-C for the evaluation of self-concept in boys or more ethnically diverse populations are encouraged to verify a similar pattern of error correlations and  $\tau$ -equivalence in their own samples before relying on estimates yielded here.

### **Conclusion**

The results of this study provide additional support for the claim that Cronbach's  $\alpha$  may be inappropriate for the evaluation of the internal consistency of some instruments. The present findings also highlight the importance of verifying the appropriateness of  $\alpha$ , *prior* to its estimation. CFA methods for estimating internal consistency are encouraged as a more durable alternative to  $\alpha$  (see Brown, 2006 for a non-technical explanation). Further, these new methods are unlikely to add meaningful pressure on current research programs, as validation of most new instruments is likely to require confirmation of its factor structure anyway. Finally, results suggest that researchers should be cautious when interpreting previously generated Cronbach's  $\alpha$  estimates, especially in cases where correlated errors have not been ruled out and  $\tau$ -equivalence has not been confirmed.

## References

- Boker S. M., Neale M. C., Maes H. H, Wilde M. J, Spiegel, M., Timothy R. Brick T. R., Spies J., et al.,(2011) OpenMx: An Open Source Extended Structural Equation Modeling Framework. *Psychometrika*, 76, 306-317
- Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, 71, 425–440.
- Brown, T. A. (2006) *Confirmatory factor analysis for applied research*. New York, NY: Guilford Press
- Boivin, M., Vitario, F., & Gagnon, C. (1992). A reassessment of the Self-Perception Profile for Children: Factor structure, reliability, and convergent validity of a French version among second through sixth grade children. *International Journal of Behavioral Development*, 15, 275- 290.
- Cairns, E., McWhirter, L., Duffy, U., & Barry, R. (1990). The stability of self-concept in late adolescence: Gender and situational effects. *Personality and Individual Differences*, 11, 937–944.
- Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. Newbury Park, CA: Sage.
- Coopersmith, S. (1967). *The antecedents of self-esteem*. San Francisco, CA: W. H. Freeman.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297-334.
- Cortina, J. M. (1993).What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, 78, 98–104.
- Cross, S. E. & Madson, L. (1997). Models of the self: Self-construals and gender. *Psychological Bulletin*, 122, 5-37

- Eapen, V., Naqvi, A., & Al-Dhaheri, A. S. (2003). Cross-cultural validation of Harter's Self-Perception Profile for Children in United Arab Emirates. *Annals of Saudi Medicine, 20*, 8-11
- Egberink, I. J. L. & Meijer, R. R. (2011). An item response theory analysis of Harter's Self-Perception Profile for Children or why strong clinical scales should be distrusted. *Assessment, 18*, 201-212
- Eiser, C., Eiser, J. R., & Havermans, T. (1995). The measurement of self-esteem: Practical implications and theoretical considerations. *Personality and Individual Differences, 18*, 429-432.
- Eriksson, M., Nordqvist, T., & Rasmussen, F. (2008). Associations between parents' and 12-year-old children's sport and vigorous activity: The role of self-esteem and athletic competence. *Journal of Physical Activity and Health, 5*, 359-373
- Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability: What they are and how to use them. *Educational and Psychological Measurement, 66*, 930-944.
- Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10*, 255-282.
- Harter, S. (1983) *The Perceived Competence Scale for Children*. Denver, CO: University of Denver.
- Harter, S. (1985) *The Self-Perception Profile for Children*. Denver, CO: University of Denver.
- Harter, S. (1993) Causes and consequences of low self-esteem in children and adolescents. In R. F. Baumeister (ed.), *Self-esteem: The puzzle of low self-regard*. New York, NY: Plenum Press.

- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55.
- James, W. (1890). *Principles of Psychology Vol. 1*. New York, NY: Henry Holt & Company
- Jöreskog, K. G., & Sörbom, D. (1996a). *LISREL 8: User's reference guide*. Chicago: Scientific Software International.
- Kapp-Simon, K. A., Simon, D. J., & Kristovich, S. (1992). Self-perception, social skills, adjustment and inhibition in young adolescents with craniofacial abnormalities. *Cleft Palate-Craniofacial Journal, 29*, 352-256.
- Kline, T. (2005). *Psychological testing: A practical approach to design and evaluation*. London, UK: Sage Publications.
- Kuder, G. F., & Richardson, M. W. (1937). The theory of estimation of test reliability. *Psychometrika, 2*, 151–160.
- Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The Sources of Four Commonly Reported Cutoff Criteria: What Did They Really Say? *Organizational Research Methods, 9*, 202-220
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, PA: Addison-Wesley.
- Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25*, 107–135.
- Muris, P., Meesters, C. & Fijen, P. (2003). The Self-Perception Profile for Children: Further evidence for its factor structure, reliability, and validity. *Personality and Individual Differences, 35*, 1791–1802.

- Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.
- Podsakoff, P. M., MacKenzie, S. M., Lee, J. Y., & Podsakoff, N. P. (2003). Common Method Biases in Behavioral Research: A Critical Review of the Literature and Recommended Remedies. *Journal of Applied Psychology, 88*, 879-903
- Raykov, T. (1997). Scale reliability, Cronbach's coefficient alpha, and violations of essential tau-equivalence with fixed congeneric components. *Multivariate Behavioral Research, 32*, 329-353.
- Raykov, T. (2001). Estimation of congeneric scale reliability using covariance structure analysis with nonlinear constraints. *British Journal of Mathematical and Statistical Psychology, 54*, 315-323.
- Raykov, T. (2002b). Analytic estimation of standard error and confidence interval for scale reliability. *Multivariate Behavioral Research, 37*, 89-103.
- Ridgers, N. D., Fazey, D. M. A., & Fairclough, S. J. (2007). Perceptions of athletic competence and fear of negative evaluation during physical education. *British Journal of Educational Psychology, 77*, 339-349.
- Schuman, B. (1999). Psychometric properties of the Self-Perception Profile for Children in a biracial cohort of adolescent girls: The NHLBI Growth and Health Study. *Journal of Personality Assessment, 73*, 260-275.
- Sciberras, E., Efron, D., & Iser, A. (2011). The child's experience of ADHD. *Journal of Attention Disorders, 15*, 321-328.
- Shevlin, M., Adamson, G., & Katrina, C. (2003). The Self-Perception Profile for Children (SPPC): A multiple-indicator, multiple-wave analysis using LISREL. *Personality and Individual Differences 35*, 1993-2005.
- Stewart, K., Roberts, M. C., & Kim, K. L. (2009). The psychometric properties of the



Harter Self-Perception Profile for Children with at-risk African American females.

*Journal of Child and Family Studies, 19, 326-333.*

Van den Bergh, B. R. H. & Van Ranst, N. (1998). Self-concept in children: equivalence of measurement and structure across gender and grade of Harter's Self-Perception Profile for Children. *Journal of Personality Assessment, 70, 564-582.*

Van Dongen, J. E. W. M, Koot, H. M., & Verhulst, F. C. (1993). Cross-cultural validation of Harter's Self-Perception Profile for Children in a Dutch sample. *Educational and Psychological Measurement 53, 739-753.*

Worth Gavin, D. A. & Herry, Y. (1996). The French Self-Perception Profile for Children: Score validity and reliability. *Educational and Psychological Measurement, 56, 678-700.*

**Appendix A: Tables****Table 1.** PA items with method bias potential

Item	Option 1	Option 2
10	Some kids are happy with their height and weight.	Other kids wish their height or weight were different.
16	Some kids wish their body was different	Other kids like their body the way it is
22	Some kids wish their physical appearance (how they look) was different	Other kids like their physical appearance the way it is

**Table 2.** Item pairs with plausible error covariances

Sub-scale	Item A	Item B	Modification Index	Substantive Justification
PA	10	16	13.70	Similar Content: Choice for different body Self Consistent Responding: nested content
PA	16	22	10.44	Similar Content: Choice for different body Self Consistent Responding: nested content
PA	10	22	9.75	Similar Content: Choice for different body Self Consistent Responding: nested content
BC	11	23	10.55	Similar Content: doing right thing/ punishment interpreted similarly
BC	17	35	10.35	Ambiguity: vague behavioral markers
SA	2	8	3.89	Self Consistent Responding: one item is consequence of other
SA	2	32	10.25	Similar Content: number of friends
SA	14	26	9.06	Similar Content: desire for more friends
SC	1	19	12.39	Self Consistent Responding: one item is consequence of other
GS	12	36	3.77	Ambiguity: vague behavioral markers
GS	24	30	3.75	Similar Content: choice for different behavior

**Table 3.** Nested models evaluating assumptions of Cronbach's  $\alpha$  for five primary SPP-C sub-scales

Model	Constraint Imposed	$X^2$ Difference	$df$ Difference	$p$
1	SC factor loadings constrained to equality	9.86	5	.079
2	AC factor loadings constrained to equality	15.74 *	5	.008
3	PA factor loadings constrained to equality	2.96	5	.706
4	BC factor loadings constrained to equality	8.78	5	.118
5	SA factor loadings constrained to equality	1.22	5	.943
6	SC Scale: Cov( Item 01, Item 19) = 0	13.17 *	1	<.001
7	PA Scale: Cov( Item 10, Item 16) = 0	20.91 *	1	<.001
8	PA Scale: Cov( Item 10, Item 22) = 0	8.80 *	1	.003
9	PA Scale: Cov( Item 16, Item 22) = 0	18.24 *	1	<.001
10	BC Scale: Cov( Item 17, Item 35) = 0	8.19 *	1	.004
11	BC Scale: Cov( Item 11, Item 23) = 0	4.12 *	1	.042
12	SA Scale: Cov( Item 02, Item 08) = 0	3.13	1	.077
13	SA Scale: Cov( Item 02, Item 32) = 0	9.78 *	1	.002

**Table 4.** Nested models evaluating assumptions of Cronbach's  $\alpha$  for GS Subscale

Model	Constraints Imposed	$X^2$ Difference	$df$ Difference	$p$
1	GS factor loadings constrained to equality	0.79	5	.978
2	GS Scale: Cov( Item 24, Item 30) = 0	4.76 *	1	.029
3	GS Scale: Cov( Item 12, Item 36) = 0	3.54	1	.060

*Note.* Both the least restrictive model ( $X^2(7) = 3.82$ ,  $p < .799$ ,  $RMSEA = 0.000$ ,  $TLI = 1.000$ ,  $CFI = 1.000$ ) and the final model ( $X^2(13) = 8.17$ ,  $p < .832$ ,  $RMSEA = 0.000$ ,  $TLI = 1.000$ ,  $CFI = 1.000$ ) resulted in acceptable fit.

**Table 5.** Reliability Estimates and Discrepancies by SPP-C Sub-scale

Sub-Scale	$\alpha$ - CFA	$\alpha$	CFA	CFA LCI	CFA UCI
GS	.013	.791	.778	.749	.807
AC	-.007	.832	.839	.792	.887
SA	-.018	.771	.790	.757	.822
PA	.136	.817	.681	.627	.735
BC	.019	.846	.827	.802	.853
SC	-.030	.791	.821	.798	.848

Appendix B: Figures

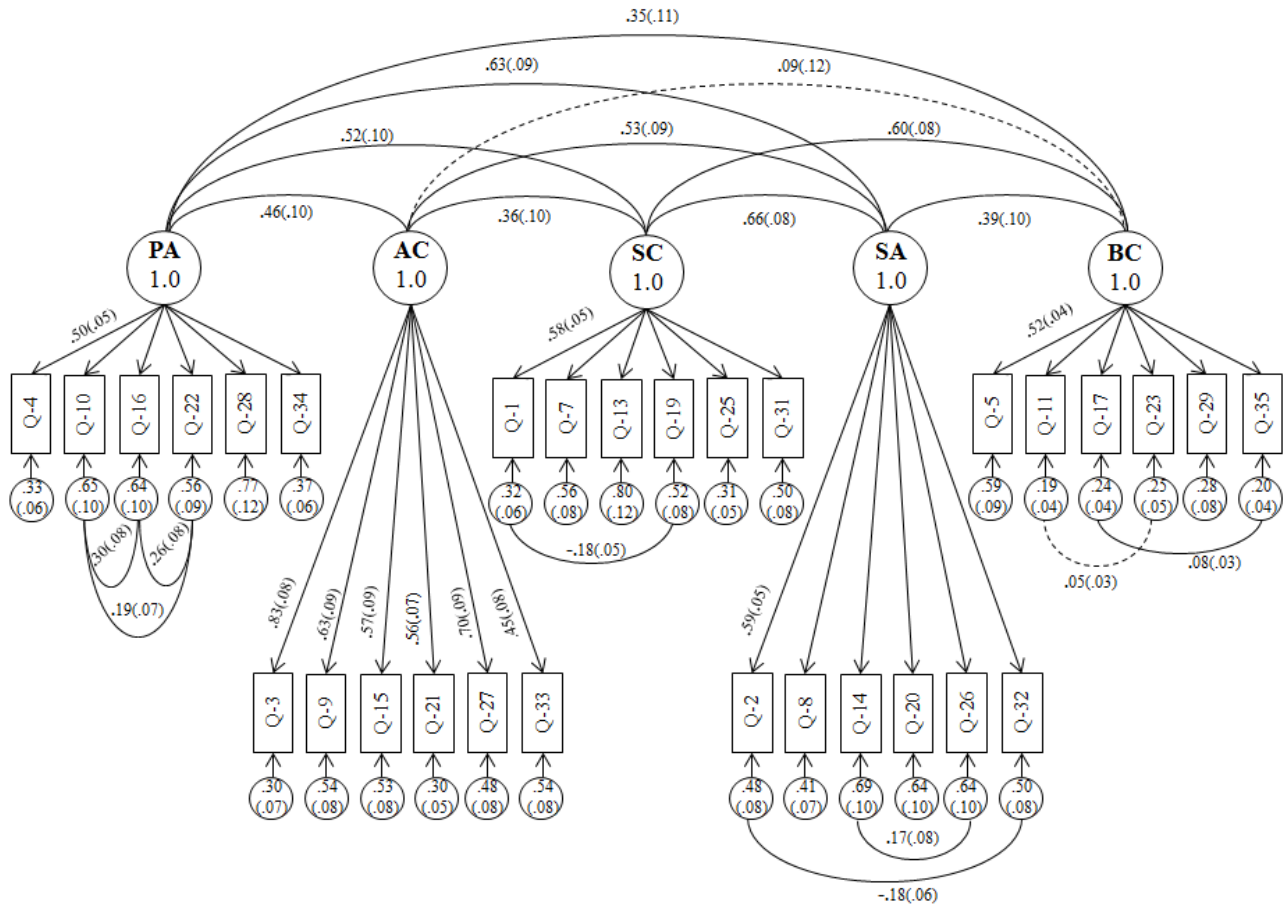


Figure 1. Path diagram relating five domain-specific SPP-C latent constructs to sub-scale items

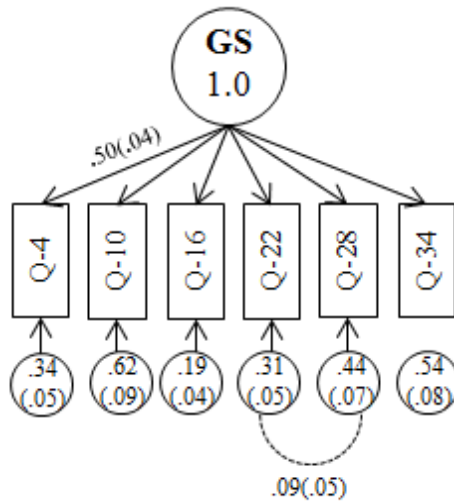


Figure 2. Path diagram relating GS latent construct to sub-scale items

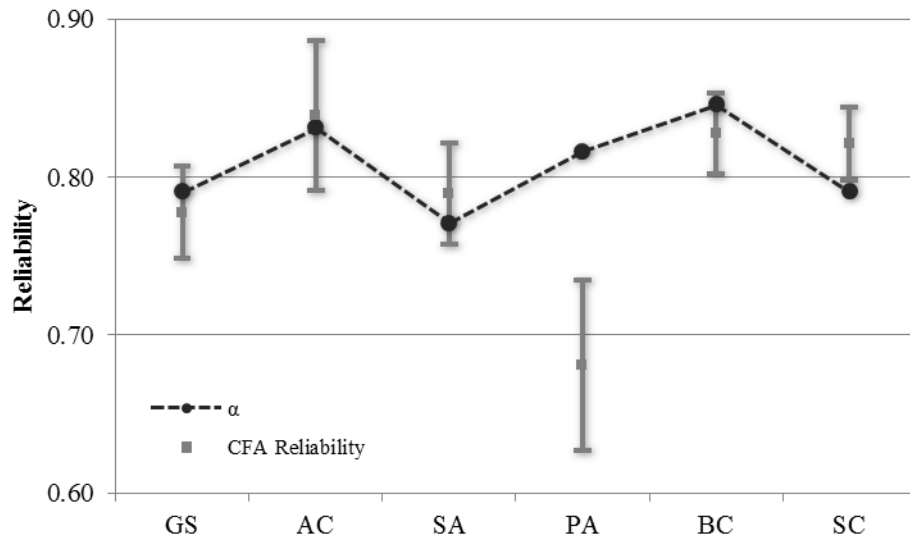


Figure 3. CFA Reliability and Cronbach's  $\alpha$  Estimates