Rethinking construct reliability within latent variable systems

Article · January 2001

CITATIONS
99

READS
3,195

2 authors, including:

Gregory R. Hancock
University of Maryland, College Park
159 PUBLICATIONS **7,986** CITATIONS

SEE PROFILE

10 Rethinking construct reliability within latent variable systems

Gregory R. Hancock & Ralph O. Mueller

Abstract

Pioneering work by Karl Jöreskog in the 1960s formalized normal theory confirmatory factor analysis (CFA), allowing researchers to specify a priori and subsequently test the number of factors underlying a set of observed variables as well as the observed measures' loadings onto each factor. Given Jöreskog's framework for assessing the plausibility of factors' measurement models, the practical question remains of how to assess the quality of a given factor. While a hypothesized model might well account for the observed (co)variances, the magnitude of the relations between a factor and its measured indicators might still call into question the reliability of that factor vis-à-vis the quality of its indicator variables. The usefulness of Cronbach's α and related CFA-based reliability measures is limited to assessing composite scales formed from a construct's indicators, rather than assessing reliability of the latent construct itself as reflected by those indicators. Here, we present a measure of construct reliability specifically for latent variable systems and show its computation and testing using the LISREL software package. In addition to its algebraic expression, we present the coefficient (1) as the squared correlation between a latent construct and the optimum linear composite formed by its indicators and (2) as a degree of attenuation relating the squared latent correlation between two constructs to the squared observed (simple, multiple, or canonical) correlation between the sets of observed indicators. Unlike other construct reliability measures, the current coefficient is never less than the best indicator's reliability. This is consistent with our position that a factor inferred from multiple indicator variables should never be worse (i.e., less reliable) than the best single indicator alone; additional indicators merely serve to enhance the construct in a manner commensurate with their own ability to reflect that construct.

10.1 Introduction

For most of the 20th century, largely exploratory factor analytic methods were employed by researchers in an attempt to inform theory as well as practice throughout the social and behavioral sciences. Those exploring their data might have used the emerged factor solution to advance an initial theory about the variables' latent underpinnings, while those with more defined expectations would use the exploratory factor analysis techniques in an attempt to validate or invalidate a specific theory about the latent structure. Given that the analytical methods were still exploratory in nature, however, the extent of the (in)validation was quite approximate, culminating at best in a highly subjective and tentative theoretical (dis)confirmation.

This changed in the late 1960s, when ground-breaking work by Karl Jöreskog (*e.g.*, 1966, 1967) formally articulated confirmatory factor analysis (CFA), the application of normal theory maximum likelihood estimation to factor models with specific theoretical latent structures. Such structures could include the *a priori* specification of the number of factors, their orthogonality or obliquity, and which variables had zero and nonzero relations with those factors. Most critical was the provision of a formal statistical test of the fit between the observed pattern of relations among the measured variables and those implied by the theorized factor model, thereby facilitating the disconfirmation or tentative confirmation of a hypothesized factor measurement model. Soon after, Jöreskog and others put forth the more general framework for the integration of measured and latent variables into complex causal networks, serving as the foundation for structural equation modeling (SEM) (*e.g.*, Bollen, 1989; Byrne, 1998; Hayduk, 1987; Mueller, 1996).

Given that a framework for assessing the plausibility of factors' measurement models had been established, the practical question remained as to the assessment of the quality of any given factor. While a factor model may well account for the observed (co)variances, the magnitude of the relations between a factor and its measured indicators might still call into question the reliability of that latent factor *vis-à-vis* the quality of its indicator variables. Several strategies for gauging reliability within a SEM paradigm have been offered and may be regarded as sequentially less restrictive variations derived from classical test theory (for a review of

classical test theory within a SEM framework, see Bollen, 1989, pp. 206–222). Miller (1995), for example, offered a simple SEM strategy using LISREL (Jöreskog & Sörbom, 1996a) for computing Cronbach's α (Cronbach, 1951) specifically for a scale that is a sum of the construct's indicator variables. Given that a scale is ultimately desired, the initial assumption is that all indicators have comparable units (*e.g.*, seven-point rating scales). In his computation, Miller further assumed essential τ-equivalence, that is, equal loadings on a single common factor and unique variances that are composed only of error. Algebraically, for a p-indicator construct, Miller's reliability equals

$$p^2\mathrm{Var}(T) \Big/ \left[p^2\mathrm{Var}(T) + p\mathrm{Var}(E)\right] \tag{10.1}$$

where $\mathrm{Var}(T)$ is the variance of the construct underlying all indicators and $\mathrm{Var}(E)$ is the equality-constrained (*i.e.*, approximate average) error variance associated with all indicators. Expression 10.1 may also be presented in the following more cumbersome manner,

$$\left(\sum_{i=1}^{p} 1\right)^2 \mathrm{Var}(T) \Big/ \left[\left(\sum_{i=1}^{p} 1\right)^2 \mathrm{Var}(T) + p\mathrm{Var}(E)\right] , \tag{10.2}$$

which will be didactically useful below. With regard to an equally weighted composite of observed indicator scores, Expressions 10.1 and 10.2 both are a ratio of the variance explainable in the composite by the underlying construct to the total variance of that indicator composite.

As Miller (1995) noted, when essential τ-equivalence fails and the measures are only congeneric (*i.e.*, when loadings on a single common factor are unequal), the above measure yields a biased underestimate of the scale reliability. Raykov (1997a), seeking to remediate this limitation, suggested a SEM approach to determine scale reliability that was more flexible in its accommodation of more general congeneric composites. Raykov's strategy, reminiscent of Drasgow & Miller's (1982) *fidelity coefficient*, modeled (indirectly) the squared correlation between the latent construct underlying a set of potentially varied-loading indicators and any desired composite formed of those indicators. The reliability associated with the unit-weighted composite is equal to

$$\left(\sum_{i=1}^{p}\lambda_i\right)^2 \mathrm{Var}(T) \Big/ \left[\left(\sum_{i=1}^{p}\lambda_i\right)^2 \mathrm{Var}(T) + \sum_{i=1}^{p}\mathrm{Var}(E_i)\right] , \qquad (10.3)$$

where $\mathrm{Var}(T)$ is the variance of the common factor underlying the indicators, and λ_i and $\mathrm{Var}(E_i)$ are the i-th indicator's unstandardized loading and residual variance, respectively. Expression 10.3, which also appears in Dillon & Goldstein (1984, p. 480, eq. 12.6–3) and as ω in McDonald (1985, p. 217, eq. 7.3.8), represents a ratio of the variance explained by the construct in a loading-weighted composite of indicators to the total variance of that composite. Thus, the primary difference between the Miller and Raykov approaches is the latter's flexibility to accommodate a congeneric latent system, as seen by the sum of loadings (Expression 10.3) rather than the sum of 1s (Expression 10.2). A secondary, and largely irrelevant, difference is that a sum of error variances is used directly rather than p times a single error variance constrained equal across indicators.

Note that Expression 10.3 may also be presented in the following manner,

$$\left[\sum_{i=1}^{p}\ell_i\mathrm{Var}(X_i)^{1/2}\right]^2 \Big/ \left(\left[\sum_{i=1}^{p}\ell_i\mathrm{Var}(X_i)^{1/2}\right]^2 + \sum_{i=1}^{p}(1-\ell_i^2)\mathrm{Var}(X_i)\right) , \qquad (10.4)$$

where ℓ_i and $\mathrm{Var}(X_i)^{1/2}$ are the i-th indicator's standardized loading and standard deviation, respectively. Expression 10.4 follows from Expression 10.3 because $\lambda_i = \ell_i\mathrm{Var}(X_i)^{1/2}/\mathrm{Var}(T_i)^{1/2}$, and is particularly illustrative as it shows the impact of the variables' variability (*i.e.*, standard deviation) on the measure of reliability. Specifically, while comparable standardized loadings may exist, a loading may be weighted more or less heavily depending on its variability. When indicators are comparably-scaled items forming a summated instrument, one would expect only mild variations in variables' variability. This expectation is an assumption for the Miller and Raykov strategies, explicitly designed for comparably-scaled items forming a summated instrument. However, these reliability measures' applicability to the more general CFA model where a construct is

indicated by entirely different measures with different units of measurement (*e.g.*, SAT scores, GPA, school attendance) is limited, and unintended.

A more general, unit-free approach was offered earlier by Fornell & Larcker (1981). Drawing on work by Werts, Linn, & Jöreskog (1974), Fornell & Larcker suggested a "reliability of the construct" (p. 45) measure (*RC*) that can be expressed as

$$\left(\sum_{i=1}^{p} \ell_i\right)^2 \Big/ \left[\left(\sum_{i=1}^{p} \ell_i\right)^2 + \sum_{i=1}^{p}(1-\ell_i^2)\right] . \qquad (10.5)$$

Note that this is precisely the Raykov measure in Expression 10.4 when the variances of all observed variables are the same. Effectively, this *RC* measure standardizes all variables and then computes Raykov's measure. Thus, *RC* is a ratio of the variance explained by the construct in a standardized loading-weighted composite of standardized indicators to the total variance of that composite.

In summary, the three strategies presented above are based on the formation of a composite of measured variables. While variations exist in the restrictivity of the nature of the true and measured composites, the approaches are grounded in classical test theory notions of composite reliability. Certainly, the reliability of composites is interesting and useful in much of social science research, and the previous measures provide some assessment of the internal consistency or interrelatedness (Green, Lissitz, & Mulaik, 1977) of a set of scale items that could be used in forming a composite. However, within SEM, unless the researcher intends to create a summated scale, a measure of the interrelatedness of a factor's indicators does not provide a direct assessment of the reliability of the construct itself as reflected in the current sample's data on the chosen indicators. In fact, in a latent variable system, the construct is perfectly reliable; instability from sample to sample only exists in the indicators' ability to reflect that latent construct.

Thus, a schism exists between the previous composite strategies and the fact that measures of reliability are desired within a completely latent variable system, which leads to some troubling properties when applying the previous measures to latent constructs. First, why should a construct's reflection be considered less reliable if it has some indicators loading with opposite sign? The quality of such a variable's indication of

the underlying construct is in no way impacted by the sign of its relation with that construct, yet any reliability coefficient assuming a positively weighted composite will be unnecessarily impoverished by an indicator loading of opposite sign. And while all negatively loading variables could be reflected around their mean so as to load positively, a desirable measure of construct reliability[1] should not require such reflection.

Second, barring additional specific factors leading to error covariances between particular indicator residuals, why should including an additional indicator ever detract from overall construct reliability? Either the variable has something (however small) to contribute to the construct's definition or it does not. Consider the extreme example of a four-indicator construct with standardized loadings .7, .7, .7, and .0. The fourth variable has no covariation with the other three, positive or negative, and hence simply does not inform the construct inferred from the other indicators. As such, the construct does not differ from that inferred using only the first three variables (*i.e.*, where standardized loadings are .7, .7, .7). Yet, the inclusion of the uninformative indicator adversely impacts reliability measures in the above modeling schemes (*e.g.*, $RC = .635$ and .742 with and without the fourth indicator, respectively) even though that indicator is not used in the construct's inference. As the reader can verify, in this example any fourth indicator with standardized loading below .476 will reduce the *RC* value.

Furthermore, why should a measure of construct reliability ever be less than that of its single best indicator? If the best indicator is a good one, using that variable alone already provides for fair reliability; employing a latent variable approach should not be worse than using the best variable alone. Moreover, any additional indicators should merely serve to enhance the construct in a manner commensurate with their own ability to reflect that construct. And if a poor indicator is chosen, certainly personal and/or theoretical consternation might be warranted. However, no penalty results in the construct's definition, and thus a measure of construct reliability should likewise not be adversely affected.

In the current chapter, we introduce a coefficient of construct reliability specifically for latent variable systems that overcomes the above limitations. The proposed measure is developed in relation to familiar relia-

[1]From this point on the term *construct reliability* will be used to refer to the stability of the construct's reflection in a given sample's data on the chosen indicators.

bility contexts, its properties are presented, and a bootstrapping strategy is offered for deriving a confidence interval around a sample estimate in order to facilitate estimation and hypothesis testing.

10.2 Development of a measure of construct reliability

Given the above justification for a better measure of construct reliability, we start by examining the role of reliability within familiar and more elementary contexts where the detection of a true population relation is desired. Whether that true relation is a population mean difference on a latent variable (as indicated by a measured proxy) or a regression relation between two latent constructs estimated by observed variables, the relation is increasingly obscured when measurement error is present. From an estimation perspective, unreliability can serve to dampen an approximation of an effect size, whether of a mean difference or regression relation. The purpose of reliability adjustments (and a purpose of latent variable methods in general) is to mitigate such a dampening, paring away the "noise" and making the "signal" as apparent as possible.

In the context of hypothesis testing of measured variable means, Cohen (1988; p. 536) noted that the standardized effect size $ES = (ES^*)(r_{YY})^{1/2}$, where ES^* is the error-free (*i.e.*, latent) standardized effect size measure and r_{YY} is the reliability of a single measured variable Y (with r_{YY} assumed to be homogeneous across groups). In squared form, Cohen's expression yields $ES^2 = (ES^*)^2(r_{YY})$, indicating that the reliability expresses how much dampening the true standardized effect size experiences as a result of measurement error in its manifest reflection. By implication then, the power to detect a true (latent) mean difference is attenuated by the reliability of the associated measured variable.

In the context of structured means models, Hancock (2000; in press) demonstrated something quite parallel to $(ES^*)^2(r_{YY})$. Under multivariate normality, the noncentrality parameter λ_0 for the χ^2 distribution associated with the test of latent mean differences may be expressed as $\lambda_0 = (N-1)f^2H$, where N is the total sample size across groups, f^2 is a squared measure of standardized effect size for J latent means, and the coefficient H (assumed to be homogeneous across groups) is a function

of individual variable reliabilities (*i.e.*, squared standardized loadings).[2] Similar to the effect of $r_{YY'}$ on $(ES^*)^2$ above, the coefficient H reflects the degree of dampening experienced by the true standardized effect size f^2 as a result of measurement error in the construct's manifest reflections. Thus, H functions precisely as a reliability estimate, but across all measured indicators of a single latent construct.

For a population $H = \mathbf{L}'\mathbf{P}^{-1}\mathbf{L}$, where $\mathbf{L}$ is a vector of p indicators' standardized loadings on a single construct, and $\mathbf{P}$ is the population correlation matrix relating the indicators. As shown by Hancock (2000; in press), H can be expressed as

$$H = \sum_{i=1}^{p} \left[\ell_i^2/(1-\ell_i^2)\right] \Big/ \left(1 + \sum_{i=1}^{p} \left[\ell_i^2/(1-\ell_i^2)\right]\right) \tag{10.6}$$

where, as before, ℓ_i is the standardized loading of the i-th indicator variable on a single latent construct. When $\mathbf{L} \neq \mathbf{0}$ (*i.e.*, when at least one standardized loading is nonzero), H can be further simplified to

$$H = 1 \Big/ \left[1 + \left(1 \Big/ \sum_{i=1}^{p} \left[\ell_i^2 \Big/ \left(1-\ell_i^2\right)\right]\right)\right] ; \tag{10.7}$$

or, for maximum clarity,

$$H = 1 \Big/ \left[1 + \frac{1}{\frac{\ell_1^2}{(1-\ell_1^2)} + \cdots + \frac{\ell_p^2}{(1-\ell_p^2)}}\right] . \tag{10.8}$$

Note that the term $\ell_i^2/(1-\ell_i^2)$ represents a ratio of the proportion of variance in the i-th variable explained by the construct (*i.e.*, the variable's reliability) to that proportion unexplained, thereby making H an aggregate function of such information across the construct's p indicator variables.

In fact, as has been recognized elsewhere (*e.g.*, Fornell & Larcker, 1981, p. 46), the quantity represented by H equals the population squared multiple correlation, P^2, from regressing the construct on its indicators, that is, the proportion of variability in the construct explainable by its own

[2] This is not the same as the H used by Bollen (1989, p. 215), which is a sum of indicator variables.

indicator variables. Thus, whereas Raykov's method yields a squared correlation between the latent construct and the unit-weighted linear composite formed from the measured indicators, H is the squared correlation between the latent construct and the *optimum* linear composite formed from the measured indicators (see discussion of Scenario C in Fig. 10.3 below).[3]

To appreciate the generality of H, consider three correlational scenarios depicted for populations in Fig. 10.1–10.3. For simplicity, all path coefficients shown represent the true standardized values in the population, and thus any model under-identification may be ignored. Scenario A shows the simplest correlational latent model, a correlation between two single-indicator constructs. The population correlation between the two measured variables ρ_{21} decomposes into $\ell_1 f_{21} \ell_2$; rearranging as $f_{21} = \rho_{21} / \left(\sqrt{\ell_1^2}\sqrt{\ell_2^2}\right)$ yields the familiar correlation correction for attenuation (*e.g.*, Nunnally & Bernstein, 1994, p. 257) using reliabilities ℓ^2. Squaring the first relation gives $\rho_{21}^2 = \ell_1^2 f_{21}^2 \ell_2^2$; this expands into the more unwieldy representation

$$\rho_{21}^2 = \left\{ 1 \Big/ \left[1 + \frac{1}{\left[\ell_1^2 / (1 - \ell_1^2)\right]} \right] \right\} f_{21}^2 \left\{ 1 \Big/ \left[1 + \frac{1}{\left[\ell_2^2 / (1 - \ell_2^2)\right]} \right] \right\} , \quad (10.9)$$

which is more simply expressed as $\rho_{21}^2 = H_1 f_{21}^2 H_2$ where H_1 and H_2 are the coefficients (*e.g.*, from Expression 10.7) for ξ_1 and ξ_2, respectively. Thus, the reliability associated with a single-indicator construct is the simplest case of coefficient H, which serves as a degree of attenuation relating the squared latent correlation to the squared observed correlation.

Scenario B expands on Scenario A, allowing ξ_2 to have multiple indicators (*e.g.*, four in Fig. 10.2). The squared multiple correlation relating X_1 to X_2 through X_5 in the population may be expressed as $P_{1.2345}^2 = \boldsymbol{\beta}' \boldsymbol{\rho}$ (*e.g.*, Pedhazur, 1997, p. 153, eq. 6.17), where $\boldsymbol{\beta}$ is the vector of standardized population regression weights from the regression of X_1 on X_2 through X_5 and $\boldsymbol{\rho}$ is the vector of population correlations relating X_1

[3]This optimum linear composite also forms the basis of the regression approach to generating factor scores (see, *e.g.*, Mulaik, 1972).

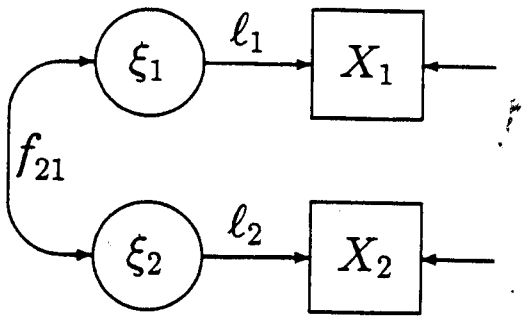

Figure 10.1 **Latent correlation scenario A**

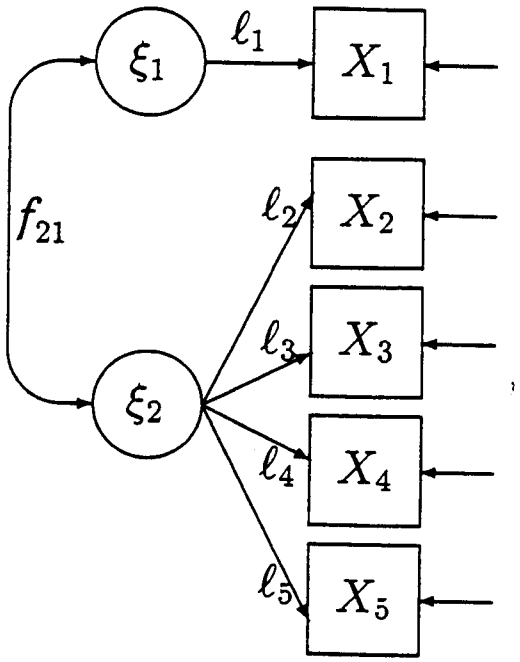

Figure 10.2 **Latent correlation scenario B**

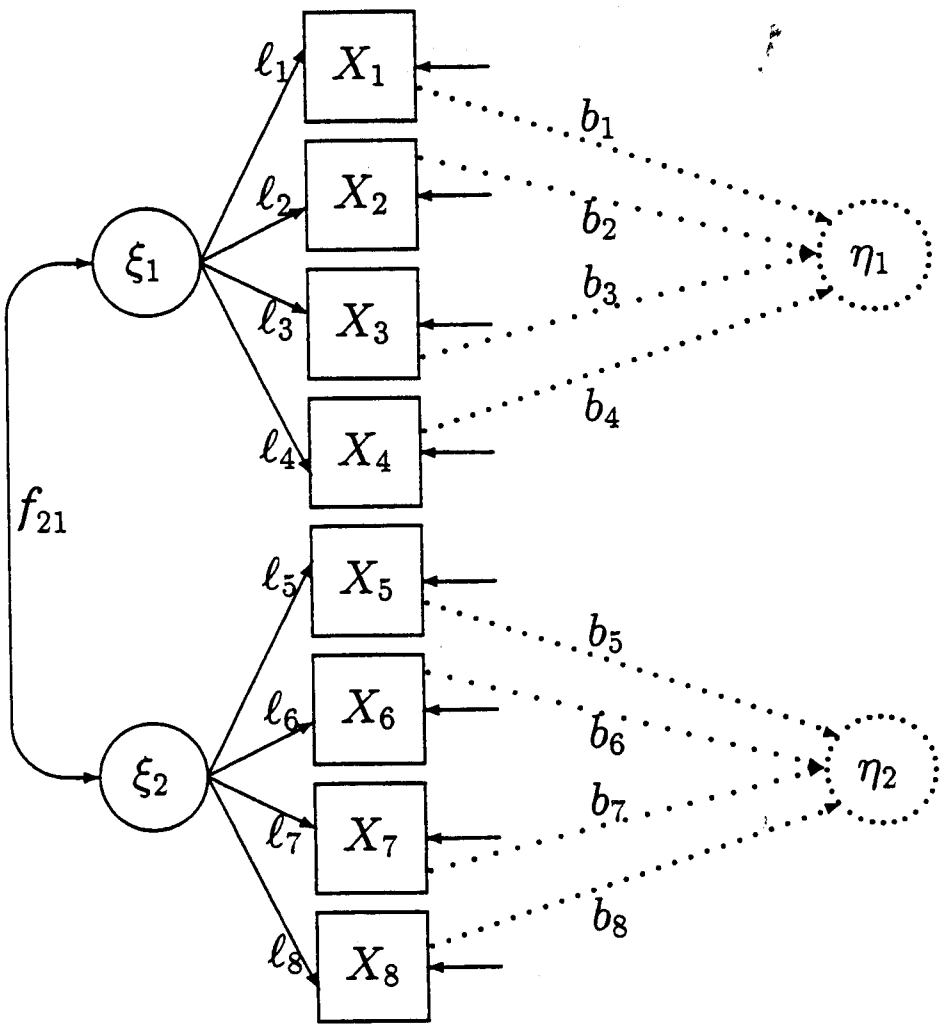

Figure 10.3 Latent correlation scenario C

to X_2 through X_5. Because $\beta = \mathbf{P}^{-1}\rho$ (*e.g.*, Pedhazur, 1997, p. 152, eq. 6.15), where $\mathbf{P}$ is the population matrix of correlations among the indicators of ξ_2, substitution (and symmetry of $\mathbf{P}^{-1}$) yields $P^2_{1.2345} = (\mathbf{P}^{-1}\rho)'\rho = \rho'\mathbf{P}^{-1}\rho$. By implication of the Scenario B population model, $\rho = \ell_1 f_{21}\mathbf{L}$ where $\mathbf{L}$ is the vector of standardized population loadings for ξ_2. Thus, substitution yields $P^2_{1.2345} = (\ell_1 f_{21}\mathbf{L})'\mathbf{P}^{-1}(\ell_1 f_{21}\mathbf{L})$; clearing parentheses and rearranging, $P^2_{1.2345} = (\ell_1^2) f^2_{21}(\mathbf{L}'\mathbf{P}^{-1}\mathbf{L})$, or more simply $P^2_{1.2345} = H_1 f^2_{21} H_2$ where H_1 and H_2 are the coefficients for ξ_1 and ξ_2, respectively. Again, we see that H serves as a degree of attenuation relating the squared latent correlation to the squared observed (now multiple) correlation.

Finally, Scenario C is the most general of the population correlational models, depicting two correlated exogenous constructs each with multiple indicators (*e.g.*, four in Fig. 10.3). New to this scenario are two endogenous constructs, η_1 and η_2, that are standardized weighted linear combinations of the indicators of the exogenous constructs ξ_1 and ξ_2, respectively. Specifically, $\eta_1 = \mathbf{b}_1'\mathbf{x}_1$, where $\mathbf{b}_1' = [b_1, b_2, b_3, b_4]$ and $\mathbf{x}_1 = [X_1, X_2, X_3, X_4]'$, and $\eta_2 = \mathbf{b}_2'\mathbf{x}_2$, where $\mathbf{b}_2' = [b_5, b_6, b_7, b_8]$ and $\mathbf{x}_2 = [X_5, X_6, X_7, X_8]'$. Here, the constructs η_1 and η_2 are meant to represent the canonical variates (synthetic variables) formed by choosing $\mathbf{b}_1$ and $\mathbf{b}_2$ so as to maximize the variates' (canonical) correlation, P_C. Using path tracing, this maximum correlation between η_1 and η_2 can be decomposed into $P_C = (\mathbf{b}_1'\mathbf{L}_1)(f_{21})(\mathbf{L}_2'\mathbf{b}_2)$, where $\mathbf{L}_1$ and $\mathbf{L}_2$ are the exogenous constructs' standardized loading vectors. The quantities in parentheses are scalars that represent the correlations between ξ_1 and η_1, between ξ_1 and ξ_2, and between ξ_2 and η_2, respectively. Squaring the above equation yields the squared population canonical correlation, $P^2_C = [(\mathbf{b}_1'\mathbf{L}_1)(f_{21})(\mathbf{L}_2'\mathbf{b}_2)]^2 = (\mathbf{b}_1'\mathbf{L}_1)^2(f_{21})^2(\mathbf{L}_2'\mathbf{b}_2)^2$. Maximizing the latter requires the squared correlation between ξ_1 and η_1 and between ξ_2 and η_2 each to be at its maximum. Thus, $\mathbf{b}_1$ and $\mathbf{b}_2$ must be chosen to yield composites η_1 and η_2 that are optimal reflections of (*i.e.*, have the greatest squared correlations with) ξ_1 and ξ_2, respectively. As mentioned previously, the greatest squared correlation for a construct and a linear combination of its indicators is coefficient H; that is, a vector $\mathbf{b}$ should be chosen so that $(\mathbf{b}'\mathbf{L})^2 = (\mathbf{L}'\mathbf{b})^2 = H$. Hence, the maximum squared correlation between η_1 and η_2 in the population – the squared canonical correlation – is achieved by choosing $\mathbf{b}_1$ and $\mathbf{b}_2$ such

that $P_C^2 = (\mathbf{b}_1'\mathbf{L}_1)^2(f_{21})^2(\mathbf{L}_2'\mathbf{b}_2)^2 = H_1 f_{21}^2 H_2$. Thus, the discussions of scenarios A, B, and C explain how coefficient H serves as a degree of attenuation relating the squared latent correlation to the squared observed correlation, be it simple, multiple, or canonical.

10.3 Characteristics of coefficient H

Given that H is seen to behave in a manner consistent with prior univariate notions of reliability, a fuller understanding of its characteristics is warranted. First, because loadings are squared in the computation of H, the sign of a loading cannot impact the assessment of construct reliability. Second, the maximum value for H is 1, occurring when a single standardized loading is 1 or -1 (if two standardized loadings are perfect, this implies two completely colinear variables and thus $\mathbf{P}$ would be singular). The minimum value for H is 0, occurring only if all standardized loadings are 0. However, while this may be true in the population, such an estimate would be unexpected for sample data as loadings will still randomly deviate from 0 in the presence of true zero loadings.

Third, because $\ell_i^2/(1-\ell_i^2) \geq 0$, readers can easily demonstrate that any additional indicator variable(s) will never decrease the value of the coefficient H (in a single-construct scenario with no error covariances). For example, for factors with $p = 3, 4, 5, 6$, and 7 indicators all with $\ell_i = .70$, values of H would be .7424, .7935, .8277, .8522, and .8706, respectively. Furthermore, by implication $H \geq \max(\ell_i^2)$ for $i = 1$ to p, meaning that H will never be smaller than the reliability (ℓ^2) of the best indicator variable. As an example, for three standardized population loadings of .90, .50, and .50, the value of H is .8314 (just over $.90^2 = .8100$).

10.3.1 Two numerical examples

To relate H back to the approaches described by Miller (1995), Raykov (1997a), and Fornell & Larcker (1981), consider as a first example the hypothetical four-variable population covariance matrix in Table 10.1, presented with the associated correlations and standard deviations. For a four-indicator construct with construct variance fixed to 1, the unstandardized population loadings for X_1, X_2, X_3, and X_4 are 12, 15, 14, and 9, respectively; corresponding standardized loadings are .30, .50, .70, and

.90. The reliability using Miller's LISREL strategy is computed to be .509, while Raykov's method accommodating varied loadings yields a slightly higher value (as expected) of .516. Fornell & Larcker's *RC*, which may be computed directly from the standardized loadings above, is .709. This is considerably higher than the previous two estimates because, as illustrated in Expression 10.4, they weight standardized loadings by the standard deviations of the measured variables. In the current example, the higher standardized loadings happen to be associated with much less variable indicators, and *vice versa*, leading to a relatively depressed reliability coefficient for the Miller and Raykov strategies. Fornell & Larcker's *RC* ignores the variables' standard deviations, thus allowing the higher standardized loadings to have more impact on the reliability coefficient. However, even the *RC* measure depicts a reliability that is less than that of the best indicator alone (X_4), $.90^2 = .81$. This is not the case with coefficient H, which may be computed with the above standardized loadings using Expression 10.8. The resulting value is $H = .850$, the highest of all measures (as expected) because the indicators are able to form a reflection of the construct better than a simple or loading-weighted sum. The optimal composite of indicators can explain 85% of the variability in the latent construct, thus indicating the construct to be more reliable than other measures would have led us to believe.

Table 10.1 Example matrices for four-indicator construct

Covariance matrix					*Correlation matrix*			
X_1	X_2	X_3	X_4		X_1	X_2	X_3	X_4
1600					1.00			
180	900				.15	1.00		
168	210	400			.21	.35	1.00	
108	135	126	100		.27	.45	.63	1.00
				SDs	40	30	20	10

As a second example, the same computations may be applied to the Hasselrot & Lernberg (1980) data in file EX71.RAW accompanying the LISREL software package (Jöreskog & Sörbom, 1996a). Data are from 200

Swedish ninth-graders who were asked about the importance of four topics believed to indicate an Equality construct and of four topics believed to indicate a Morality construct.[4] Focusing just on the Equality construct (with variance fixed to 1), the four indicators HUMRGHTS (human rights), EQUALCON (equality conditions for all people), EQUALVAL (equal value of all people), and EUTHANAS (euthanasia) have unstandardized loadings of .320, .569, .641, and .640, respectively; corresponding standardized loadings are .421, .712, .712, and .797. Also, standard deviations for the variables are .76, .80, .90, and .80, respectively. The data yield a reliability estimate of .752 using Miller's strategy, .767 using Raykov's approach, .763 using Fornell & Larcker's *RC* measure, and a sample estimate of the population coefficient H of $\widehat{H} = .801$. Notice that while $\widehat{H}$ is still higher in value, all estimates are quite close: most standardized loadings are strong, and the variables have fairly similar standard deviations. As indicators' loading quality and variability start to diverge, so too do the different reliability estimates, as seen in the previous example.

10.3.2 Magnitude of coefficient H and variables' relative contributions

The inevitable question is "How big should coefficient H be for a given construct?" or, alternatively, "How much of a construct should be explainable by the chosen indicators?" Within a single measured variable context, Nunnally & Bernstein (1994, p. 265) recommended minimum reliability levels of .70 or .80. For the *RC* measure, Hair, Anderson, Tatham, & Black (1998, p. 612) mentioned .70 or higher as desirable. If one accepts these values as reasonable goals for coefficient H, the implications for standardized loadings are as follows. In general, as the reader can verify (*e.g.*, using Expression 10.6), to achieve a target level of H, at least t out of p indicators must have standardized loadings of

$$\ell_i = [H/(H - Ht + t)]^{1/2} . \qquad (10.10)$$

Thus, a value of $H = .70$ would be achieved if any single indicator ($t = 1$) has $\ell_i = .837$, if two indicators ($t = 2$) have $\ell_i = .734$, if three

[4]Although the data appear in the LISREL manual to illustrate procedures for ordinal data, we are treating the data as intervally scaled for the purposes of this and the bootstrapping example later in the chapter.

indicators ($t = 3$) have $\ell_i = .661$, if four indicators ($t = 4$) have $\ell_i = .607$, and in general if t out of p indicators have $\ell_i = [7/(3t + 7)]^{1/2}$. For more stringent construct reliability, to achieve a value of $H = .80$ would require any single indicator ($t = 1$) to have $\ell_i = .894$, two indicators ($t = 2$) to have $\ell_i = .816$, three indicators ($t = 3$) to have $\ell_i = .756$, four indicators ($t = 4$) to have $\ell_i = .707$, and in general t out of p indicators to have $\ell_i = [4/(t + 4)]^{1/2}$. Situations might exist where lower reliability is tolerable, particularly if no other indicators exist in the construct's domain. But the consequence of having low construct reliability is a reduced certainty in the magnitude of relations observed among constructs in the given sample. This is because stronger loadings tend to fluctuate less from sample to sample, leading to more stability in the inferred constructs and hence more stable estimates of the latent relations within the model.

The above discussion of desired magnitude of H also raises the related question of the contribution of each indicator to defining the construct. Drawing from the previous notion that H is a population squared multiple correlation (P^2) relating the observed indicators to the latent construct, one can consider evaluating the utility of a variable in terms of ΔP^2, or ΔH. For population data, an indicator's ΔH value could be determined directly by analyzing the model with and without an indicator (assuming enough indicators to avoid under-identification, $p \geq 4$) and computing the difference between resulting H values. Alternatively, ΔH may be computed using the standardized loadings from the model with all indicators, where ΔH is the difference between H values with and without an indicator's standardized loading. For a properly specified model in a population, standardized loadings will be identical with and without a given indicator, and thus both approaches would yield identical values of ΔH for any indicator. For example, for the population data presented in Table 10.1, computing ΔH both ways yields the same values: .002, .008, .025, and .268 for X_1, X_2, X_3, and X_4, respectively, when each is individually removed as a manifest variable. This indicates that any of the first three variables provide little additional information in defining the construct.

Deriving an estimate of ΔH based on sample data, $\Delta \widehat{H}$, the manner in which it is done does indeed matter. In the properly specified population scenario above, it is the absence of residual covariance that allows the standardized loadings to remain unchanged when an indicator is re-

moved. However, when some unmodeled residual covariance is present, as is usually the case in fitting sample data to an over-identified model, solutions for loading parameters adjust positively or negatively in an attempt to accommodate that covariation. Then, the removal of an indicator leads to the removal of a set of up to $p - 1$ nonzero residual covariances that no longer exert a competing force on the loading estimation, resulting in loading changes and thus possibly unexpected changes in $\widehat{H}$. In fact, if any nonzero residual covariance is present, sample loadings might increase when an indicator is deleted, possibly leading to an actual *increase* in $\widehat{H}$. If the residual covariance causing the loading increase is indicative of some real, unmodeled specific factor(s), the single-factor model was already misspecified from the start; data-model fit indices and/or modification indices might alert the researcher to this misspecification. Given a proper model, the residual covariance is just random covariability, and hence spurious loading increases and decreases may result, as might a spurious increase in $\widehat{H}$.

For the Hasselrot & Lernberg (1980) sample data involving the Equality construct, recall that the standardized loadings are .421, .712, .712, and .797, yielding $\widehat{H} = .801$. The values of $\Delta\widehat{H}$ estimated from this full model are .009, .051, .051, and .106, respectively, while the corresponding values of $\Delta\widehat{H}$ estimated when dropping single indicators from the model are .004, $-.029$, .024, and $-.067$. The two negative values indicate that $\widehat{H}$ increases when dropping either of those observed variables in this sample, implying counterintuitively that dropping the strongest loading indicator will increase construct reliability. We recommend using the first set of $\Delta\widehat{H}$ values in the context of the full model, implying that the first indicator (HUMRGHTS) offers virtually nothing toward establishing a reliable construct; the second and third indicators (EQUALCON and EQUALVAL) also offer relatively little compared to the fourth indicator (EUTHANAS). Our recommendation to use $\Delta\widehat{H}$ values based on the full complement of estimated standardized loadings proceeds under the assumption that the factor model is indeed properly specified (in the absence of data-model fit information to the contrary). This strategy will make all $\Delta\widehat{H}$ values exist in the context of the measurement model chosen *a priori*, and reduce the exploratory temptation to drop indicators based on possibly random fluctuations in data.

10.3.3 Inference regarding coefficient H

Similar to the approach presented by Raykov (1997a), the bootstrapping feature of PRELIS (Jöreskog & Sörbom, 1996b) can be used to estimate the standard error of $\widehat{H}$ and to derive an empirical confidence interval for the population coefficient H. Alternative approaches exist, such as a possible extension of work by Neale & Miller (1997) for generating a confidence interval for (Miller's estimate of) Cronbach's α, or drawing upon distribution theory for R^2 values (see Algina, 1999) because $\widehat{H}$ is equivalent to an R^2 for the regression of the construct on its own indicators. However, these strategies rely upon strong distributional assumptions, the violation of which could compromise the accuracy of the resulting standard errors and confidence intervals. Thus, a distribution-free, bootstrapping alternative seems most prudent.

The Appendix (see p. 214) contains a set of three annotated input files that were used for the Equality construct from the Hasselrot & Lernberg (1980) raw data, reflecting the three-step bootstrapping process outlined in Jöreskog & Sörbom (1996b, pp. 184–188). First, 1,000 bootstrapped variance/covariance matrices (with replacement) for the observed variables HUMRGHTS, EQUALCON, EQUALVAL, and EUTHANAS (treating data as continuous for this example) are generated using PRELIS. Second, bootstrapped $\widehat{H}$ values, $\widehat{H}_{BS}$, obtained from each of the 1,000 bootstrapped variance/covariance matrices are computed using LISREL. The input file presented in the Appendix utilizes LISREL's AP (additional parameters) option (Jöreskog & Sörbom, 1996a, p. 347–348) to compute the standardized loadings for the four observed variables on the underlying Equality construct by emulating a one factor CFA model and expressing the correlations among the observed variables, ϕ_{ij}, as functions of the unknown standardized loading parameters. A latent endogenous variable, η (Eta), was created for the sole purpose of storing $\widehat{H}_{BS}$ as the structural coefficient, GA(1,1). Finally, the Appendix shows a PRELIS input file that produces the mean, standard deviation, and other descriptive statistics of the 1,000 saved $\widehat{H}_{BS}$ values.

For the current data set, recall that $\widehat{H} = .801$. For the distribution of the 1,000 $\widehat{H}_{BS}$ values, the mean is 0.811 and the standard deviation is .037. The standardized skewness ($-.952$; $p = .341$) and standardized kurtosis (5.514; $p < .001$) together indicate that the empirical $\widehat{H}_{BS}$

sampling distribution deviates from normality ($\chi^2 = 31.307$; $p < .001$). Thus, rather than constructing confidence intervals around $\widehat{H} = .801$ assuming normality of $\widehat{H}_{BS}$, we determined the empirical 90% and 95% confidence intervals using appropriate percentiles within the bootstrapped sampling distribution: $CI_{90} = (.750, .865)$ and $CI_{95} = (.734, .879)$. These may be used to facilitate hypothesis testing. For example, using a two-tailed test at either the .10 or .05 level, the null hypothesis of $H = .70$ would be rejected in favor of higher construct reliability; the null hypothesis of $H = .80$ would be retained as both confidence intervals contain this hypothesized value.

10.4 Conclusion

From within Jöreskog's CFA framework (1966, 1967), we sought to address a topic of importance not only to CFA, but also to the general structural equation modeling context: How does a researcher assess the quality of a latent construct that is indicated by several observed variables? In this chapter we built on previous conceptualizations of construct reliability to rethink its possible assessment within latent variable systems. Whereas approaches such as those offered by Miller (1995), Raykov (1997a), or Fornell & Larcker (1981) are outgrowths of classical test theory (conceptualizing reliability based on the formation of a composite of measured variables), we operationalized construct reliability by focusing on the reliability of the construct itself as reflected by the indicators chosen. Coefficient H describes the relation between the latent construct and its measured indicators in a manner more consistent with discussions of reliability in other research contexts (*e.g.*, disattenuating bivariate correlations and detecting mean differences). Unlike other measures of construct (or composite) reliability, we demonstrated that coefficient H is unaffected by the sign of indicators' loadings, drawing information from all indicators in a manner commensurate with their ability to reflect the construct.

The proposed coefficient H, and its associated ΔH, are based on the assumption of a properly specified latent variable model that may include multiple factors with both covariance and regression relations among

those factors. In addition, variables may cross-load on multiple factors,[5] and error covariances can be included in the model where theoretically justifiable. It is our hope that SEM researchers and programmers will further explore the utility and limitations of coefficient H and its associated ΔH measure.

10.5 Appendix: PRELIS and LISREL syntax for bootstrapping $\widehat{H}$

10.5.1 Step 1

This PRELIS input file generates 1,000 variance/covariance matrices for the Equality factor by bootstrapping samples of size $n = 200$ from the Hasselrot & Lernberg (1980) data (LISREL example file EX71.RAW). Here, the observed variables are treated as continuous, as the CO ALL command indicates. The four indicators of the Morality factor are excluded from analysis with the select cases/delete variables (SD) command. The chosen number of bootstrap samples (BS) is 1,000 with a sampling fraction (SF) of 100%; that is, subsamples of size equal to the parent sample size are drawn with replacement. All generated variance/covariance matrices (CM) are stored in the file H_BS.CMB.

```
DA NI=8
LA
HUMRGHTS EQUALCON RACEPROB EQUALVAL EUTHANAS CRIMEPUN CONSCOBJ GUILT
RA FI=c:\lisrel83\LS8EX\EX71.RAW
CO ALL
SD RACEPROB CRIMEPUN CONSCOBJ GUILT
OU MA=CM BS=1000 SF=100 BM=H_BS.CMB
```

10.5.2 Step 2

This LISREL input file computes 1,000 (as specified with the RP option) bootstrap estimates of coefficient H from the variance/covariance matrices (CM) saved in H_BS.CMB during Step 1.

To compute coefficient H using Expression 10.8, the standardized factor loadings of the four observed variables on the factor EQUALITY are

[5]Note that for a variable indicating multiple factors, it will have a value of $\Delta\widehat{H}$ for each construct served.

needed. This is accomplished by using LISREL's additional parameters (AP) option. By constraining the standardized covariances of the observed variables (PH=ST) to appropriate products of the additional parameters, LISREL solves the resulting system of equations so that PAR(1) through PAR(4) contain the values of the estimated standardized loadings. Because LISREL does not currently support the use of parentheses in equality constraints, PAR(5) through PAR(14) were created for the sole purpose of storing intermediate results. A latent endogenous variable without indicator variables or prediction error was created (NE=1; TE=ZE; PS=ZE) so that $\widehat{H}_{BS}$ may be stored in the structural gamma coefficient. All 1,000 $\widehat{H}_{BS}$ values were stored as gamma coefficients in the file H_BS.PV. To minimize the size of the list output file, the XO option on the OU command (only output for the first replication is given) was specified.

```
DA NI=4 NO=200 RP=1000
LA
HUMRGHTS EQUALCON EQUALVAL EUTHANAS
CM=H_BS.CMB
MO NX=4 NK=4 NE=1 TD=ZE TE=ZE LX=DI,FR PH=ST GA=FU,FI PS=ZE AP=14

!  Estimate the standardized loadings (PAR1 - PAR4)
CO PH(2,1)=PAR(1)*PAR(2)
CO PH(3,1)=PAR(1)*PAR(3)
CO PH(4,1)=PAR(1)*PAR(4)
CO PH(3,2)=PAR(2)*PAR(3)
CO PH(4,2)=PAR(2)*PAR(4)
CO PH(4,3)=PAR(3)*PAR(4)

!  Compute estimate of H according to Expression 8
CO PAR(5)=1-PAR(1)^2
CO PAR(6)=1-PAR(2)^2
CO PAR(7)=1-PAR(3)^2
CO PAR(8)=1-PAR(4)^2
CO PAR(9)=PAR(1)^2*PAR(5)^-1
CO PAR(10)=PAR(2)^2*PAR(6)^-1
CO PAR(11)=PAR(3)^2*PAR(7)^-1
CO PAR(12)=PAR(4)^2*PAR(8)^-1
CO PAR(13)=PAR(9)+PAR(10)+PAR(11)+PAR(12)
CO PAR(14)=1+PAR(13)^-1

CO GA(1,1)=PAR(14)^-1   !Compute and store estimate of H in GA(1,1)

ST 0.5 PAR(1) - PAR(14) !Provide starting values for the additional parameters
OU GA=H_BS.PV XO        !and designate output
```

10.5.3 Step 3

This PRELIS program provides descriptive statistics for the 1,000 $\widehat{H}_{BS}$ values stored in H_BS.PV during Step 2. The $\widehat{H}_{BS}$ values are treated as continuous (CO ALL).

```
DA NI=1
LA
H_HAT_BS
RA=H_BS.PV
CO All
OU
```