Running Head: SCALED DIFFERENCE CHI-SQUARE TESTING

Structural Equation Modeling (in press)

Principles and Practice of Scaled Difference Chi-Square Testing

Fred B. Bryant

LOYOLA UNIVERSITY CHICAGO

&

Albert Satorra

UNIVERSITAT POMPEU FABRA

Contact information:

Professor Fred B. Bryant
Department of Psychology
Loyola University Chicago
1032 W. Sheridan Road
Chicago, IL 60660
Tel: 773-508-3033
fbryant@luc.edu

Professor Albert Satorra
Department of Economics
Universitat Pompeu Fabra
Ramon Trias Fargas 25-27
08005-Barcelona, Spain
Tel: 34- 93-542-17-58
albert.satorra@upf.edu

Abstract

We highlight critical conceptual and statistical issues and how to resolve them in conducting Satorra-Bentler (SB) scaled difference chi-square tests. Concerning the original (Satorra & Bentler, 2001) and new (Satorra & Bentler, 2010) scaled difference tests, a fundamental difference exists in how to compute properly a model's scaling correction factor (c), depending on the particular SEM software used. Because of how LISREL defines the SB scaled chi-square, LISREL users should compute c for each model by dividing the model's NTWLS chi-square by its SB chi-square, to recover c accurately with both tests. EQS and Mplus users, in contrast, should divide the model's ML chi-square by its SB chi-square to recover c. Because ML estimation does not minimize the NTWLS chi-square, however, it can produce a negative difference in nested NTWLS chi-square values. Thus, we recommend the standard practice of testing the scaled difference in ML chi-square values for models M_1 and M_0 (after properly recovering c for each model), to avoid an inadmissible test-numerator. We illustrate the difference in computations across software programs for the original and new scaled tests and provide LISREL, EQS, and Mplus syntax in both single- and multiple-group form for specifying the model M_{10} that is involved in the new test.

Principles and Practice of Scaled Difference Chi-Square Testing

1. Introduction

The Scaled Chi-Square Statistic

Structural equation modeling (SEM) relies heavily on goodness-of-fit chi-square statistics to assess the adequacy of hypothesized models as representations of observed relationships. However, multivariate nonnormality is known to inflate overall goodness-of-fit test statistics (Kaplan, 2000). Accordingly, Satorra and Bentler (1988, 1994) developed a set of corrected normal-theory test statistics that adjust the goodness-of-fit chi-square for bias due to multivariate nonnormality. Correcting the regular chi-square value for nonnormality requires the estimation of a scaling correction factor (c), which reflects the amount of average multivariate kurtosis distorting the test statistic in the data being analyzed. One divides the goodness-of-fit chi-square value for the model by the scaling correction factor to obtain the so-called Satorra-Bentler (SB) scaled chi-square.

The SB scaled chi-square has performed well in Monte Carlo simulation studies (e.g., Chou, Bentler & Satorra, 1991; Curran, West & Finch, 1996; Hu, Bentler & Kano, 1992) and has become well accepted in the SEM literature. The SB chi-square is currently available in three of the four software packages most often used to conduct SEM—EQS (Bentler, 1995), LISREL (Jöreskog & Sörbom, 1996a), and Mplus (Muthén & Muthén, 2007) report the SB chi-square, whereas AMOS (Arbuckle, 2006, 2007) does not.

Testing Differences in Nested Chi-Square Values

Among the most versatile and commonly-used strategies for hypothesis-testing in SEM is the likelihood-ratio test, also known as the difference chi-square test (Bollen, 1989), with which researchers contrast the goodness-of-fit chi-square value of a less restrictive, baseline model (M_1) with the goodness-of-fit chi-square value of a more restrictive, nested comparison model (M_0).

One typically obtains the more restrictive comparison model (M_0) by placing constraints, such as fixed values, invariance restrictions, or equality constraints, on particular parameters in the baseline model (M_1). Because the difference in goodness-of-fit chi-square values for two nested models is itself distributed as chi-square (see Neyman & Pearson, 1928; Steiger, Shapiro, & Browne, 1985), researchers can subtract the chi-square value of the baseline model (M_1) from the chi-square value of the nested comparison model (M_0) and use the resulting difference in chi-square values (with accompanying difference in degrees of freedom) to test the hypothesis that the constraints imposed on the baseline model significantly worsen model fit.

If, on the one hand, this difference chi-square is statistically significant, then one rejects the null hypothesis and concludes that the baseline model fits the data better than the nested comparison model. If, on the other hand, the difference chi-square is nonsignificant, then one fails to reject the null hypothesis and concludes that the nested comparison model fits the data just as well as does the baseline model. As with the overall goodness-of-fit chi-square value itself, however, the validity of statistical conclusions drawn from the difference in nested chi-square values is suspect under conditions of multivariate nonnormality.

2. The Original Scaled Difference Chi-Square Test

Whereas the traditional difference chi-square test allows researchers to compare directly the fit of nested models when using standard goodness-of-fit chi-square values, this is not the case when using SB scaled chi-square values. In particular, the difference in SB scaled chi-square values for nested models does not correspond to a chi-square distribution (Satorra, 2000). For this reason, simply subtracting the SB chi-square value for the less restrictive, baseline model (M_1) from the SB chi-square value for the more restrictive, comparison model (M_0) yields an invalid statistic for testing hypotheses about differences in model fit.

To overcome this limitation, Satorra (2000) derived a formula for testing the difference in nested SB chi-square values, to permit scaled difference chi-square testing. However, because this formula uses statistical information not readily available in conventional SEM software, it is an impractical approach for most applied researchers. Accordingly, Satorra and Bentler (2001) developed a simpler, asymptotically equivalent procedure for scaled difference chi-square testing that is easily implemented using the scaled and unscaled chi-square values and the degrees of freedom for the two models contrasted.

This "original" scaled difference test (Satorra & Bentler, 2001) requires the user to estimate the baseline model (M_1) and comparison model (M_0), in order to obtain the standard goodness-of-fit chi-square value and SB chi-square value for each model. Because the SB chi-square is defined as a standard goodness-of-fit chi-square value divided by a scaling correction factor (Satorra & Bentler, 1988, 1994), dividing the standard goodness-of-fit chi-square by the SB chi-square for a particular model allows the user to "recover" the scaling correction factor (c) for that model, for use in scaled difference testing. With the original scaled difference test, the user computes the scaling factors for models M_1 and M_0 by dividing the standard chi-square value by the SB chi-square value for each model.

Nonequivalent Definitions of the SB Chi-Square across Software Programs

It is a little known fact, however, that not all SEM software packages use the same goodness-of-fit chi-square to define the SB scaled chi-square. For this reason, the particular "standard" goodness-of-fit chi-square that one should use to recover each model's scaling correction factor depends on the specific software that one uses. In particular, EQS 6 (Bentler, 1995, p. 218) and Mplus 6 (Muthén & Muthén, 2007, Appendix 4, pp. 357-358) base the SB chi-square value on a rescaling of the maximum-likelihood (ML) minimum-fit function chi-square. LISREL 8, in contrast, bases the SB chi-square value on a rescaling of the normal-theory

weighted least-squares (NTWLS) chi-square (Jöreskog, Sörbom, du Toit, & du Toit, 1999, Appendix A, pp. 191-202). For this reason, EQS and Mplus users must use the ML chi-square to recover each model's scaling factor, whereas LISREL users must use the NTWLS chi-square. Because it does not estimate the Satorra-Bentler chi-square, AMOS does not permit users to implement the scaled difference chi-square test (though this feature may be added in the future).

A Computational Error LISREL Users Commonly Make in Scaled Difference Testing

Because researchers traditionally conduct chi-square difference testing by contrasting the ML chi-square values for models M_0 and M_1, LISREL users might naturally assume that the scaling correction factors for the models are also based on ML chi-square values. As noted above, however, the proper chi-square to use to recover the scaling correction factor for each model depends on how the particular software program has defined the SB scaled chi-square.

For present purposes, we delineate the following terms:

T_1 = maximum-likelihood (ML) chi-square test statistic (which Jöreskog et al., 1999, called *C*1)

T_2 = normal theory weighted least-squares (NTWLS) chi-square test statistic (which Jöreskog et al., 1999, called *C*2)

T_3 = Satorra-Bentler (SB) scaled chi-square test statistic (which Jöreskog et al., 1999, called *C*3)

m = (*df* for model M_0) – (*df* for model M_1)

Then for LISREL (see Jöreskog et al., 1999, Appendix A, pp. 191-202), the scaling correction factor for a given model is:

$$c = T_2/T_3$$

But for EQS (see Bentler, 1995, p. 218) and Mplus (Muthén & Muthén, 2007, Appendix 4, pp. 357-358), the scaling correction factor for a given model is:

$$c = T_1/T_3$$

Mplus users have two options for obtaining c for a given model—they can compute the ratio of T_1/T_3, or they can take c straight from the Mplus output—both of which provide the same result. For all three software programs, the scaling correction factor (i.e., denominator) for the original Satorra-Bentler scaled difference test (c_d) is then:

$$((df \text{ for model } M_0) \text{ x } (c \text{ for } M_0) - (df \text{ for model } M_1) \text{ x } (c \text{ for } M_1))/m$$

Finally, to compute the original scaled difference chi-square test statistic (with $df = m$), one divides the difference in chi-square values for models M_0 and M_1 by this difference scaling correction factor (c_d). Because of how each software program defines the scaled chi-square, the original scaled difference test for LISREL users is:

$$((T_2 \text{ for } M_0) - (T_2 \text{ for } M_1))/(c_d)$$

Whereas the original scaled difference test for users of EQS or Mplus is:

$$((T_1 \text{ for } M_0) - (T_1 \text{ for } M_1))/(c_d)$$

Because of the lack of formal documentation, however, LISREL users are likely to make the mistake of using the ML chi-square values for models M_1 and M_0 in both the numerator and denominator of the formula for the original scaled difference test. For example, the website for Mplus provides instructions for "Chi-Square Difference Testing Using the Satorra-Bentler Scaled Chi-Square" in which users are told to divide a model's "regular chi-square value" by its SB chi-square value, in order to recover the scaling correction factor for the particular model (see http://www.statmodel.com/chidiff.shtml). Here, as we noted above, the exact meaning of "regular chi-square" depends on the particular SEM software program that one uses.

Prior work in LISREL using Satorra and Bentler's (2001) original scaled difference chi-square testing procedure has been incorrect if it has used minimum-fit function (ML) chi-square values, rather than NTWLS chi-square values, to recover the scaling correction factor for models

M_1 and M_0 in the test denominator. In addition, available macro programs that compute scaled difference chi-square values using ML and SB chi-square values alone as input (e.g., Crawford & Henry, 2004; see SBDIFF.EXE at http://www.abdn.ac.uk/~psy086/dept/psychom.htm, and online calculator at http://www.uoguelph.ca/~scolwell/difftest.html) produce correct results for EQS and Mplus users, but will produce incorrect scaling factors and test values when based on the ML and SB chi-square values reported in the LISREL output.

Thus, researchers who have analyzed their data via LISREL and have used available macro programs to conduct scaled difference testing based on LISREL results (e.g., McDonald, Hartman, & Vrana, 2008; Schaffer, Vogel, & Wei, 2006; Warren, Cepeda-Benito, Gleaves, Moreno, Rodriguez, Fernandez, Fingeret, & Pearson, 2007) have reported findings that are technically inaccurate, although scaled difference testing might well lead to the same substantive conclusions if computed accurately. We can speculate that the improper use of the ML chi-square to recover the scaling correction factors for the models contrasted via LISREL explains some of the reported instances of negative difference chi-square values when using the original scaled difference test.

A Simple Method for LISREL Users to Compute a Scaled Maximum-Likelihood Goodness-of-Fit Test Statistic

After LISREL users have recovered the scaling correction factor (c) for a given model by dividing the model's NTWLS chi-square value (T_2) by its SB chi-square value (T_3), they can obtain a scaled ML chi-square test statistic for the particular model simply by dividing the model's ML chi-square value (T_1) by c. This approach enables LISREL users to compute a scaled ML test statistic for a single model that is equivalent to the scaled values produced by EQS and Mplus, thereby obtaining a comparable form of Satorra-Bentler scaled chi-square statistic across software packages. The LISREL scaled ML test statistic for a given model—

defined as the model's ML chi-square value (T_1) divided by a scaling factor (c) recovered from the model's NTWLS chi-square value (T_2/T_3)—thus enables a meaningful comparison of scaled chi-square values across different SEM software programs. To facilitate an equivalent definition of the SB chi-square across software programs, we recommend that LISREL users routinely report SB chi-square values based on ML test statistics, just as EQS and Mplus users do.

This scaled ML goodness-of-fit test statistic (T_1/c) also enables LISREL users to obtain accurate results when using available macro programs to compute scaled difference chi-square values. Specifically, if LISREL users input T_1 as the "normal" chi-square value and T_1/c as the SB chi-square value for models M_1 and M_0, then this macro program provides accurate results for LISREL users.

A Potential Problem When Testing Scaled Differences in NTWLS Chi-Square Values

Scaling the minimum-fit function (ML) chi-square value also helps to avoid a potential problem that arises when one scales the difference in NTWLS chi-square (T_2) values for the models contrasted in difference testing. Because ML estimation does not minimize the NTWLS chi-square (T_2) the way it does the ML chi-square value (T_1), the value of T_2 for the more restrictive model M_0 can actually be smaller than the value of T_2 for the less restrictive model M_1, such that (T_2 for M_0) – (T_2 for M_1) < 0. As a consequence, it is possible that LISREL users who correctly use T_2 instead of T_1 to recover the scaling correction factors for models M_1 and M_0 may still obtain a negative scaling correction factor for the original scaled difference chi-square test, when they contrast T_2 for the two models in the numerator of the scaled difference test formula. We note that using the NTWLS chi-square value is not a problem when evaluating the goodness-of-fit of a single model. Rather, it is scaling the difference in NTWLS chi-square values for models M_1 and M_0 that can be problematic.

A colleague (Mary Johnson of James Madison University) has shared with us an empirical example that dramatically illustrates the potential pitfall of scaling differences in NTWLS chi-square values. This researcher used LISREL 8 to contrast all pairs of factor intercorrelations in a three-factor CFA model by conducting three scaled difference tests, each of which produced a negative test result when using the original formula (Satorra & Bentler, 2001), even when correctly defining each model's scaling factor as T_2/T_3. These inadmissible results occurred because the NTWLS chi-square value of model M_0 was less than the NTWLS chi-square value of model M_1 for all three tests. For each of the three model contrasts, on the other hand, dividing the difference in ML chi-square (T_1) values for models M_0 and M_1 by c_d produced a proper positive scaled difference test result for LISREL. Clearly, the best practice in SEM is to test scaled differences in ML chi-square (T_1) values, rather than scaled differences in NTWLS chi-square (T_2) values. (Note that the ML chi-square value of the more restrictive model M_0 can sometimes be smaller than the ML chi-square of the less restrictive model M_1 when the ML chi-square values for the two models are equal but the convergence criterion is too large—a "harmless" problem of numerical imprecision that can be solved by specifying a more stringent convergence criterion; see http://www.statmodel.com/discussion/messages/9/156.html?1271351726).

To reduce the likelihood of obtaining a negative value in scaled difference testing, we recommend that LISREL users test and report differences in T_1 (rather than differences in T_2) for models M_0 and M_1, just as EQS and Mplus users routinely do. To implement this modified ML version of the original scaled difference test, LISREL users should employ the following steps:

1. Recover the scaling correction factor (c) for each model by dividing its NTWLS chi-square value by its SB chi-square value (T_2/T_3).
2. Multiply the scaling correction factor (c) for each model by the model's *df*.

3. Subtract this product for model M_1 from the same product for model M_0.
4. Divide the result by m (i.e., the difference in df between models M_0 and M_1), to obtain the scaling factor for the difference test (c_d).
5. Finally, divide the difference in the ML chi-square (T_1) values of models M_0 and M_1 by the scaling factor for the difference test (c_d), with df for the scaled difference test (m) = (df for model M_0 – df for model M_1).

EQS and Mplus users should first recover c by dividing each model's ML chi-square by its SB chi-square (T_1/T_3), or Mplus users can take c directly from the model output, and should then follow the remaining steps 2-5 as outlined above.

3. Illustrating the Original Scaled Difference Chi-Square Test

To clarify these steps, we now present a worked example of the computations involved in scaled difference chi-square testing using Satorra and Bentler's (2001) original formula—first for LISREL users, and then for EQS and Mplus users. Data for these analyses consist of a sample of 803 American undergraduates (647 females, 156 males) who completed the 12-item revised version of the Life Orientation Test (LOT-R; Scheier, Carver, & Bridges, 1994), a commonly-used self-report measure of dispositional optimism. The LOT-R consists of 4 positively-worded items, 4 negatively-worded items, and 4 unscored "filler" items with which respondents indicate their extent of agreement on a 5-point scale (0 = strongly disagree, 1 = disagree, 2 = neutral, 3 = agree, 4 = strongly agree).

Previous researchers (e.g., Bryant & Cvengros, 2004; Chang, D'Zurilla, & Maydeu-Olivares, 1994) have found that a congeneric two-factor model—consisting of correlated Optimism (positively-worded items) and Pessimism (negatively-worded items) factors—provides an excellent fit to responses to the 8 scored LOT items and fits significantly better than a one-factor model, which provides a poor fit to the data. With the present data, initial single-

group CFA replicated these prior findings and indicated the oblique two-factor model provided a good fit in terms of RMSEA, SRMR, CFI, and NNFI.

For present purposes, we tested the hypothesis that dispositional optimism has more to do with positive future expectancies than with benefit-finding in the face of adversity. Specifically, we employed single-group CFA to compare the loading of LOT item 5 ("I'm always optimistic about my future") on the Optimism factor (unstandardized loading = .932) and the loading of LOT item 11 ("I'm a believer in the idea that 'every cloud has a silver lining'") on the Optimism factor (unstandardized loading = .526); and we used Satorra and Bentler's (2001) original test to assess the statistical significance of the difference in the size of these two factor loadings for the pooled sample.

We analyzed covariance matrices specifying robust maximum-likelihood estimation. Computation of the SB chi-square also requires estimation of the asymptotic covariance matrix. For LISREL 8.80, we first used PRELIS 2.0 (Jöreskog & Sörbom, 1996b, pp. 167-171) to compute and store the asymptotic covariance matrices for the 8 LOT items, for use as input files along with raw data for CFA, specifying METHOD=ML on the OUTPUT line to obtain robust ML estimation. For EQS 6.1, we analyzed raw data specifying METHOD=ML, ROBUST to obtain robust ML estimation (Bentler, 1995, pp. 46-48). For Mplus 6.1, we set ESTIMATOR=MLM on the ANALYSIS line to obtain robust estimation (Muthén & Muthén, 2007, p. 533).

Following established SEM procedures for testing differences in estimated parameters, comparing the magnitude of the factor loadings involves contrasting the goodness-of-fit chi-square values of two models: (a) a baseline model (model M_1) in which the two loadings being compared are freely estimated; and (b) a nested comparison model (model M_0) in which the two

loadings being compared are constrained to be equal. The difference in chi-square values between these two models provides an inferential test regarding the difference in factor loadings.

Specifying Models M_1 and M_0

For all three software programs, the single-group CFA syntax for model M_1 specified: (a) eight measured variables and two latent variables; (b) a pattern of factor loadings in which the loadings of the four positively-worded LOT items were declared free on the first (Optimism) factor but were fixed at zero on the second (Pessimism) factor, and the loadings of the four negatively-worded LOT items were declared free on the second (Pessimism) factor but were fixed at zero on the first (Optimism) factor (and one loading was fixed at a value of 1.0 for each factor to define the units of variance for the two latent variables); (c) a pattern of factor variances and covariance for the two latent variables in which all parameters were freely estimated; and (d) independent unique error variances for each of the eight measured variables. For LISREL, EQS, and Mplus the single-group CFA syntax for model M_0 was identical to the syntax for model M_1, except that it included an equality constraint that forced the estimated value of the two contrasted loadings (for LOT items 5 and 11) to be equal in magnitude.

Computing the Original Scaled Difference Test

Table 1 illustrates the computations involved in conducting the original scaled difference test (Satorra & Bentler, 2001) for users of LISREL, EQS, and Mplus (see table entries 1-8). For LISREL users, we include two sets of computations—one for testing scaled differences in the NTWLS (T_2) chi-square values of models M_0 and M_1 (see table entry 9); the other, for testing scaled differences in the ML (T_1) chi-square values of models M_0 and M_1 (see table entry 10). We advocate using the *latter* ML-based approach to avoid obtaining an inadmissible negative value for the numerator of the scaled difference test. For EQS and Mplus users, we include only computations for scaled ML (T_1) difference chi-square testing (table entry 10).

Inspecting the results displayed in Table 1 for the original scaled ML difference test, we see that overall the general conclusions are the same across the three software programs, although there is a noticeable difference between the results of the test when using Mplus ($\Delta\chi^2_{SB}$ = 78.1766035) versus either EQS ($\Delta\chi^2_{SB}$ = 65.3342186) or LISREL ($\Delta\chi^2_{SB}$ = 65.2382987). Indeed, two discrepancies across software programs are evident in Table 1, namely differences in the value of the scaling correction factor (c) and differences in the value of the ML chi-square (T_1). We now comment and explain these two discrepancies in turn.

The discrepancy in the value of c is not an issue of the difference test itself, but rather stems from how the three software programs compute the scaling correction factor for the goodness-of-fit test as originally presented by Satorra and Bentler (1988, 1994). The formula for c involves a normal-theory (NT) weight-matrix (W), which in turn involves a consistent estimate of the population covariance matrix (Σ). For ML estimation, EQS and LISREL base this estimate of Σ on the fitted Σ, while for generalized least-squares (GLS) estimation EQS uses the sample covariance matrix (S). Using our own software, we determined that Mplus uses S in computing W, regardless of estimation method. Supporting this conclusion, when we changed estimation method from ML to GLS and specified robust estimation, the scaling correction factors produced by EQS and Mplus agree to several decimal digits. So, in summary, the discrepancy in c seen in Table 1 arises from the use of S (in Mplus) versus fitted Σ (in EQS and LISREL) to compute the weight-matrix involved in the formula for the scaling correction factor.

Regarding the second discrepancy—that is, differences in the value of the ML chi-square (T_1)—our own computations lead us to conclude that whereas EQS and LISREL both report the minimum of the ML fitting function (when requesting ML estimation) and the minimum of the NT-GLS fitting function (when requesting GLS estimation), Mplus provides the value of the

fitting function using the multiplier n instead of the multiplier $n - 1$ that is used in EQS and LISREL. This discrepancy thus should vanish when sample size is large enough.

We note that, although the general conclusions of the scaled difference test converge across software programs, the discrepancy in the final scaled difference test chi-square is remarkable. This unexpected result demonstrates that alternative expressions that are equivalent in abstract theoretical form can in actual practice produce surprising and puzzling discrepancies. However, we anticipate that the formulae used by the alternative software programs will be equivalent asymptotically. A classic example of this phenomenon is the choice between the unbiased estimate of the population covariance matrix Σ, which is S divided by $n - 1$, versus the ML estimate of Σ, which is S divided simply by n. Both estimates are valid and in fact converge for large samples, but can yield striking discrepancies in small samples; and those discrepancies will grow larger as sample size decreases. (In fact, in this issue while Mplus computes S dividing by n, EQS and LISREL compute S dividing by $n - 1$. With our own software we determined that, for the data set considered, the difference of using n versus $n - 1$ in computing the matrix S in the Mplus calculations of the scaling correction has no noticeable effects on the final value of the test statistic. This finding contrasts with the noticeable divergence we found in the value of T_1 as a result of the different software programs using n versus $n - 1$.) The observed discrepancy in the final scaled difference test chi-square convinces us of the need to explore and better understand differences across the various SEM software programs that are available to users in producing the same statistics.

4. The New Scaled Difference Chi-Square Test

Although Satorra and Bentler's (2001) original scaled difference chi-square test has been widely used, it sometimes produces a negative scaling correction factor that leads to a negative difference in chi-square values, particularly in small samples or when the more restrictive model

(M_0) is highly incorrect. For this reason, Satorra and Bentler (2010) recently proposed an improved scaling correction procedure that precludes negative differences in chi-square values and produces results identical to those obtained when using Satorra's (2000) complex formula.

As with the original scaled difference test, the new scaled difference test (Satorra & Bentler, 2010) requires the user to estimate and obtain goodness-of-fit statistics for the baseline model (M_1) and comparison model (M_0). With the new scaled difference test, however, the user must also estimate the baseline model with the number of iterations fixed at zero, using the final parameter estimates from M_0 as starting values (termed "model M_{10}"). As with the original scaled difference test, the new test requires the user to compute the scaling correction factor (c) for model M_0 by dividing the proper chi-square value by the SB chi-square value for this model. With the new scaled difference test, the user also computes c for model M_{10} by dividing the proper chi-square value for model M_{10} by the SB chi-square value for model M_{10} (i.e., T_2/T_3 for LISREL users; T_1/T_3 for EQS and Mplus users). One then uses c for model M_{10} in place of c for model M_1, to compute the correction factor for the new scaled difference test (c_d).

To conduct the new scaled difference test, one follows the same computational steps as with the original scaled difference test, except that one replaces the scaling correction factor (c) for model M_1 in the denominator with the scaling correction factor for model M_{10}. The scaling factor for the new Satorra-Bentler scaled difference test (c_d) is thus: ((df for model M_0) x (c for M_0) – (df for model M_1) x (c for M_{10}))/m. As with the original scaled difference test, we recommend that LISREL users compute the new scaled difference test based on differences in ML (T_1) values in the numerator, to avoid situations in which (T_2 for M_0) – (T_2 for M_1) < 0. Using this latter ML-based numerator in LISREL will also promote a single uniform scaled test statistic that is comparable across SEM software programs.

Recall the empirical example we noted earlier in connection with the original scaled difference test in which a colleague used LISREL 8 to contrast pairs of factor intercorrelations by conducting three scaled difference tests, each of which produced a negative difference in NTWLS chi-square values for models M_0 and M_1. When applying the new scaled difference test, the same inadmissible results occurred, since the numerator of the scaled difference test is identical for both the original and new formulas. Although the new scaled difference test is designed to avoid an inadmissible negative test statistic, it can only do so if the chi-square value for model M_1 is less than or equal to the chi-square value for model M_0. Thus, we suggest that all SEM users, regardless of software, test the scaled difference in ML chi-square (T_1) values for models M_1 and M_0 when using either the original or new scaled difference test.

Specifying Model M_{10}

To clarify how to set up model M_{10} for the new scaled difference test, we now explain how to specify this model in LISREL, EQS, and Mplus. Because the new scaled difference test has not yet been widely disseminated, we also provide readers with examples of the LISREL, EQS, and Mplus syntax required to set up model M_{10} in both single-group CFA (see Appendix A) as well as multigroup CFA (see Appendix B). Applied users can find other descriptions of the single-group syntax for specifying model M_{10} via: (a) EQS in the Appendix of the preprint version of Satorra and Bentler (2010), which can be downloaded at http://preprints.stat.ucla.edu/539/Satorra-Bentler%20Manuscript.pdf; and (b) Mplus in Appendix A of Asparouhov and Muthén (2010), which can be downloaded at http://www.statmodel.com/examples/webnotes/webnote12.pdf.

For LISREL, EQS, and Mplus, the single-group CFA syntax for model M_{10} is identical to the syntax for model M_1, except for two modifications: (a) it includes a matrix of starting values consisting of factor loadings, factor variances and covariances, and unique error variances taken

directly from the final parameter estimates in the output for model M_0; and (b) the number of iterations is frozen at zero. Although model M_{10} has the same pattern of fixed and free elements as model M_1, note that model M_{10} fixes the parameter values in model M_1 to the final estimates for model M_0, and model M_{10} should exclude the equality constraints added to model M_0.

LISREL users can export the final estimates in model M_0 directly to separate external ASCII files for each parameter matrix using the Output command (Jöreskog & Sörbom, 1996, p. 95), and they can then specify each external file as the source of starting values for each parameter matrix in model M_{10} (Jöreskog & Sörbom, 1996, p. 84). However, LISREL always exports the matrix of final estimates for unique error variances (Theta Delta; TD) in a symmetric form, even when the TD matrix for model M_1 is specified as diagonal (e.g., TD=DI,FR). As a result, if model M_1 specifies TD as diagonal, then LISREL users who import starting values for model M_{10} from external files exported from model M_1 must change the syntax for model M_{10} to specify TD as a symmetric matrix with free diagonal elements and fixed subdiagonal elements.

A second option for LISREL users in setting up model M_{10} for single-group CFA is to manually copy and paste the final estimates from the output file for model M_0 into the syntax file for model M_{10}, and then specify these final estimates as starting values using MA commands for the Lambda-x, Phi, and Theta-Delta matrices in the CFA model. If one uses this option, then one should replace the dashed lines (i.e. "- -") that LISREL reports for fixed values of zero in the parameter matrices of the output file for model M_0 with values of 0.0 in the matrix of starting values in the syntax file for model M_{10}. We chose this second option as our means of estimating model M_{10} in LISREL. We fixed iterations at zero by specifying IT=0 in the Output command.

For single-group EQS, we obtained the starting values for model M_{10} by specifying a "retest" file (i.e., RETEST=*newfile*) in the PRINT section of the syntax file for model M_0, thereby storing the final parameter estimates of model M_0 in a separate outfile (Bentler, 1995, p.

257). We then manually copied and pasted these final estimates for model M_0 into the syntax file for model M_{10}. We fixed iterations at zero by specifying ITER=0 in the Technical section of the syntax file for model M_{10}.

To conduct the new scaled difference test, Mplus 6 includes the option OUTPUT: SVALUES that facilitates the creation of the syntax file for model M_{10} by generating syntax in the output file for model M_0 that sets starting values equal to the fitted values for model M_0. Mplus users can copy and use this syntax as the syntax file for model M_{10}. However, note that because model M_{10} should exclude the invariance constraints added to model M_0, Mplus users must delete the numbers in parentheses included in the SVALUES output for model M_0, which indicate the equality-constrained parameters added to model M_0. Although Mplus does not allow users to specify ITERATIONS=0, Mplus users can freeze iterations to estimate model M_{10} by specifying a very large convergence criterion (e.g., CONVERGENCE=100000000). Specifying the TECH5 option on the OUTPUT command prints the iteration history, thereby enabling users to inspect the Mplus output for model M_{10} to verify whether they have set the convergence criterion large enough to prevent iterations, or whether they must increase it to halt iterations. Using a large convergence criterion successfully freezes iterations at zero for LISREL (when omitting IT=0), but does not stop iterations for EQS (for which only ITER=0 freezes iterations). To help SEM users conduct the new scaled difference test, Appendix A provides the single-group LISREL, EQS, and Mplus syntax we used to estimate model M_{10}.

A Technical Anomaly in LISREL 8

In applying the new scaled difference chi-square test, we have discovered a technical problem that occurs when using LISREL 8.80 (Jöreskog & Sörbom, 1996) to estimate model M_{10}. Specifically, when contrasting two divergent parameter estimates, LISREL produces values for the SB chi-square (T_3) that are too small and values for the scaling correction factor (c) that

are too large, when freezing iterations at zero to estimate model M_{10}; but when contrasting two parameter estimates that are highly similar in magnitude, LISREL produces values for T_3 and c that are accurate. Thus, we have found that LISREL can produce a *negative* scaling correction factor for the new difference test when contrasting two parameters that are very different in value. (We have informed the distributors of LISREL about this anomaly, which they have acknowledged and will undoubtedly resolve in a future software release.)

Computing the New Scaled Difference Test

Table 1 also illustrates the computations involved in the new scaled difference test (Satorra & Bentler, 2010) for users of LISREL, EQS, and Mplus (see table entries 11-13). Because we do not advocate NTWLS (T_3) difference chi-square testing, we have included only computations for scaled ML (T_1) difference chi-square testing (table entry 13) using the new scaled test. In computing the new scaled difference test, we have omitted the results for LISREL 8.80 because of the program's technical anomaly mentioned above in the case of model M_{10}. We have also omitted the value of T_2 for model M_{10} because the NTWLS chi-square value is only relevant for scaled difference testing via LISREL in recovering the value of c for model M_{10}.

Inspecting the results displayed in Table 1 for the new scaled ML difference test, we see that the overall conclusions are the same across software programs, although there is a noticeable discrepancy between the results of the new test (see table entry 14) when using Mplus ($\Delta\chi^2_{SB}$ = 60.1363734) versus EQS ($\Delta\chi^2_{SB}$ = 53.7247488). As with the original scaled difference test, this discrepancy across software programs arises from differences in the value of the scaling correction factor (c) and differences in the value of the ML chi-square (T_1). As noted earlier, in computing the weight-matrix involved in the formulae for the scaling correction factor and the fitting function, Mplus uses S (the sample covariance matrix) whereas EQS uses Σ (the fitted estimate of the population covariance matrix)—a computational difference that is asymptotically

equivalent, but produces a somewhat smaller c and larger T_1 for the same model using Mplus versus EQS, when the null hypothesis is false and sample size decreases.

5. Conclusion

This paper makes several contributions that we feel are of importance for the practice of SEM. We have clarified how the specific methods of scaled difference testing differ fundamentally in LISREL, versus EQS or Mplus, and we have illustrated the correct procedures for recovering the scaling correction factors and implementing the original (Satorra & Bentler, 2001) and new (Satorra & Bentler, 2010) tests for both groups of software users. We have identified a mistake LISREL users are prone to make in computing the scaling correction factor for a particular model. We have highlighted specific situations in which LISREL can produce inadmissible results for either the original or new scaled difference test. And we have presented evidence supporting a uniform ML approach to scaled difference chi-square testing.

The primary purpose of this paper is to help SEM analysts implement scaled difference chi-square testing properly. Toward this goal, we have highlighted three potential pitfalls and how to avoid them, in using LISREL to implement scaled difference chi-square testing. First, because LISREL obtains the SB chi-square by scaling the NTWLS chi-square (whereas EQS and Mplus scale the ML chi-square), LISREL users who base a model's scaling correction factor on its ML chi-square value will obtain inaccurate results for both the original and new scaled difference tests. When computing the original or new scaled difference test, LISREL users can avoid this problem by using the NTWLS chi-square value rather than the ML chi-square to recover each model's scaling correction factor (i.e., $c = T_2/T_3$, *not* T_1/T_3). For users of EQS and Mplus, on the other hand, $c = T_1/T_3$, not T_2/T_3.

Contrasting NTWLS chi-squares in scaled difference testing creates another potential problem. Because maximum-likelihood estimation minimizes the ML chi-square but not

necessarily the NTWLS chi-square, it is possible for the NTWLS chi-square value of the less restrictive model (M_1) to be smaller than the NTWLS chi-square value of the more restrictive model (M_0), especially when the contrasted parameter values are highly similar and sample size is small. This circumstance will produce an inadmissible negative difference in model chi-square values (i.e., test numerator) when computing the original or new scaled difference test. This potential problem exists whenever contrasting the values of nested NTWLS chi-squares.

To reduce the likelihood of obtaining inadmissible negative values in scaled difference testing, we recommend that LISREL users test differences in ML chi-square values (instead of differences in NTWLS chi-square values), by dividing the difference in ML chi-square (T_1) values for models M_0 and M_1 by the correction factor for the scaled difference test (c_d), after first recovering each model's scaling correction factor, c, by T_2/T_3. Given that EQS and Mplus users routinely test scaled differences in ML chi-square (T_1) values, using this standard ML-approach in LISREL also offers the advantage of making the type of scaled chi-square statistic that researchers report comparable across SEM software packages.

We have also highlighted a third, temporary pitfall—only relevant until the distributors correct the software in a future release—that LISREL users face in implementing the new scaled difference test (Satorra & Bentler, 2010). Our results reveal that for certain set-ups LISREL 8.80 produces an inflated scaling correction factor for the analysis of M_{10}, which in turn can lead to an improper negative scaling correction factor for the new scaled difference test. We suggest that LISREL distributors also consider changing the program's definition of the SB chi-square from T_2/c to T_1/c (as both EQS and Mplus define it), to facilitate a single, standard ML-based scaled chi-square statistic and a uniform ML approach to scaled difference chi-square testing in SEM.

The question naturally arises as to when users should employ the new versus original scaled difference test. Asparouhov and Muthén (2010) have suggested that users adopt the new

test when the original test produces a negative statistic or when the original correction factor is very small. Given that the new test requires evaluating only one more model than the original test, our recommendation is that users routinely employ the new difference test, to be sure of avoiding a negative scaling correction factor.

A final point concerns situations in which it may be difficult or impossible to specify model M_{10}. Specifying model M_{10} should be relatively simple for the standard forms of difference testing we have described, where model M_0 represents baseline model M_1 with some restrictions added to the parameters of M_1. A technical assumption implicit in difference testing (though rarely recognized, and even difficult to assess in applications) is that the rank of the Jacobian matrix associated with model M_1 is regular (constant rank) at any point of model M_0. This assumption, which Satorra and Bentler (2010) made explicit, is required for difference testing in general, even with normally distributed data where scaling corrections are unnecessary.

This assumption may fail, however, when M_0 sets parameters of M_1 at the boundary of their permissible values (e.g., if M_1 is a two-factor CFA model, and M_0 fixes the variance of a factor to zero), thereby producing difficulties in computing the new scaling correction via M_{10}. Indeed, practitioners using either the original or new correction formula—or not using scaling corrections at all—may fail to note a rank deficiency problem in the particular difference testing considered, and may thus compute a difference test statistic that looks proper but is incorrect because it is not actually a chi-square statistic. (See Hayashi, Bentler & Yuan, 2007, for an example of a non-standard set-up that does in fact distort difference testing.) In most typical applications, such as setting regression coefficients to zero, equating loading coefficients across groups, or constraining factor covariances to be equal, this constant rank assumption holds true, and in fact it is implicitly assumed. Although comparing scaled statistics in non-standard settings is beyond the scope of the present paper, we intend to pursue this issue in further research.

References

Arbuckle, J. L. (2006). *AMOS 7.0 user's guide*. Spring House, PA: Amos Development Corporation.

Arbuckle, J. L. (2007). *AMOS 16.0 user's guide*. Chicago: SPSS, Inc.

Asparouhov, T., & Muthén, B. (2010). Computing the strictly positive Satorra-Bentler chi-square test in Mplus. (Mplus Web Notes: No. 12, July 6, 2010). See http://www.statmodel.com/examples/webnotes/webnote12.pdf.

Bentler, P. M. (1995). *EQS structural equations program manual.* Encino, CA: Multivariate Software.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Bryant, F. B., & Cvengros, J. A. (2004). Distinguishing hope and optimism: Two sides of a coin, or two separate coins? *Journal of Social and Clinical Psychology, 23,* 273-302.

Chang, E. C., D'Zurilla, T. J., & Maydeu-Olivares, A. (1994). Assessing the dimensionality of optimism and pessimism using a multimeasure approach. *Cognitive Therapy and Research, 18*, 143 -160.

Chou, C. -E., Bentler, P. M., & Satorra, A. (1991). Scaled test statistics and robust standard errors for nonnormal data in covariance structure analysis: A Monte Carlo study. *British Journal of Mathematical and Statistical Psychology, 44,* 347-357.

Crawford J.R., & Henry J.D. (2004). The Positive and Negative Affect Schedule (PANAS): Construct validity, measurement properties, and normative data in a large non-clinical sample. *British Journal of Clinical Psychology, 43,* 245–265.

Curran, R. J., West, S. G., & Finch, J. E (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods,* 1, 16-29.

Hayashi, K., Bentler, P. M., & Yuan, K.-H. (2007). On the likelihood ratio test for the number of factors in exploratory factor analysis. *Structural Equation Modeling, 14,* 505–526.

Hu, L., Bentler, P. M., & Kant, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin, 112,* 351-362.

Jöreskog, K., & Sörbom, D. (1996a). *LISREL 8: User's reference guide.* Chicago: Scientific Software International.

Jöreskog, K., & Sörbom, D. (1996b). *PRELIS 2: User's reference guide.* Chicago: Scientific Software International.

Jöreskog, K., Sörbom, D., du Toit, S., & du Toit, M. (1999). *LISREL 8: New statistical features*. Chicago: Scientific Software International.

Kaplan, D. (2000). *Structural equation modeling: Foundations and extensions*. Thousand Oaks, CA: Sage.

McDonald, S. D, Hartman, N. S, & Vrana, S. R. (2008). Trait anxiety, disgust sensitivity, and the hierarchic structure of fears. *Journal of Anxiety Disorders, 22,* 1059-1074.

Muthén, L. K., & Muthén, B. O. (2007). *Mplus user's guide*. Los Angeles: Muthén & Muthén.

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika, 20A,* 175-240.

Satorra, A. (2000). Scaled and adjusted restricted tests in multisample analysis of moment structures. In D. D. H. Heijmails, D. S. G. Pollock, & A. Satorra (Eds.), *Innovations in multivariate statistical analysis: A Festschrift for Heinz Neudecker* (pp. 233-247). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Satorra, A., & Bentler, P. M. (1988). Scaling corrections for chi-square statistics in covariance structure analysis. *ASA 1988 Proceedings of the Business and Economic Statistics, Section* (308-313). Alexandria, VA: American Statistical Association.

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 399-419). Thousand Oaks, CA: Sage.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika, 66,* 507-514.

Satorra, A., & Bentler, P. M. (2010). Ensuring positiveness of the scaled chi-square test statistic. *Psychometrika, 75,* 243-248.

Schaffer, P. A., Vogel, D. L., & Wei, M. (2006). The mediating roles of anticipated risks, anticipated benefits, and attitudes on the decision to seek professional help: An attachment perspective. *Journal of Community Psychology, 53,* 442-452.

Scheier, M. F., Carver, C. S., & Bridges, M. W. (1994). Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem): A reevaluation of the Life Orientation Test. *Journal of Personality and Social Psychology, 67*, 1063-1078.

Steiger, J. H., Shapiro, A., & Browne, M. W. (1985). On the multivariate asymptotic distribution of sequential chi-square statistics. *Psychometrika, 50*, 253−264.

Warren, C. S., Cepeda-Benito, A., Gleaves, D. H., Moreno, S., Rodriguez, S., Fernandez, M. C., Fingeret, M. C., & Pearson, C. A. (2007). English and Spanish versions of the Body Shape Questionnaire: Measurement equivalence across ethnicity and clinical status. *International Journal of Eating Disorders, 41,* 265-272.

Acknowledgments

This research was supported in part by a grant SEJ2006-13537 from the Spanish Ministry of Science and Technology (to Albert Satorra). The authors express their gratitude to Mary Johnson, for sharing the results of scaled difference tests from her master's thesis in which (T_2 for model M_0) < (T_2 for model M_1); Karen Saban, for help in discovering the anomalous scaling factor when IT=0 in LISREL 8.80; Bengt Muthén, for suggesting the use of a more stringent convergence criterion to avoid cases in which (T_1 for model M_0) < (T_1 for model M_1), and the use of a more liberal convergence criterion to freeze iterations at zero when using Mplus to estimate model M_{10}; Gerhard Mels, for clarifying LISREL options in specifying model M_{10} and exporting LISREL parameter estimates to external files; and Juan Carlos Bou (Universitat Jaume I, Castelló, Spain), for help in constructing Mplus syntax.

Table 1

Computing the Original and New Scaled Difference Tests Using LISREL 8.80, EQS 6.1, and Mplus 6.1

		Software Program		
Result	**Statistic**	**LISREL**	**EQS**	**Mplus**
1. Model M_1	*df*	19	19	19
	T_1	42.970	42.974	43.027
	T_2	43.896	- -	- -
	T_3	36.827	36.053	36.214
	c	43.896/36.827 = 1.1919516	42.974/36.053 = 1.1919674	43.07/36.214 = 1.1881317
2. Model M_0	*df*	20	20	20
	T_1	108.443	108.451	108.584
	T_2	111.455	- -	- -
	T_3	94.251	91.715	92.755
	c	111.455/94.251 = 1.1825339	108.451/91.715 = 1.1824783	108.584/92.755 = 1.1706539

Table 1 (continued)

		Software Program		
Result	**Statistic**	**LISREL**	**EQS**	**Mplus**
3. Difference in NTWLS χ^2 values	(T_2 for model M_0) – (T_2 for model M_1)	111.455 – 43.896 = 67.559	- -	- -
4. Difference in ML χ^2 values	(T_1 for model M_0) – (T_1 for model M_1)	108.443 – 42.970 = 65.473	108.451 – 42.974 = 65.477	108.584 – 43.027 = 65.557
5. Difference in *df* for models M_0 & M_1	*m*	20 – 19 = 1	20 – 19 = 1	20 – 19 = 1
6. Term 1 for scaling factor of difference test	(*df* for model M_0) x (*c* for model M_0)	20 x 1.1825339 = 23.6506780	20 x 1.1824783 = 23.6495660	20 x 1.1706539 = 23.4130780
7. Term 2 for scaling factor of *original* difference test	(*df* for model M_1) x (*c* for model M_1)	19 x 1.1919516 = 22.6470804	19 x 1.1919674 = 22.6473806	19 x 1.1881317 = 22.5745023

Table 1 (continued)

		Software Program		
Result	**Statistic**	**LISREL**	**EQS**	**Mplus**
8. Scaling factor for *original* scaled difference test (original c_d)	((*df* for model M_0) x (*c* for model M_0) – (*df* for model M_1) x (*c* for model M_1))/*m*	(23.6506780 – 22.6470804)/1 = 1.0035976	(23.6495660 – 22.6473806)/1 = 1.0021854	(23.4130780 – 22.5745023)/1 = 0.8385757
9. *Original* scaled NTWLS difference test	((T_2 for model M_0) – (T_2 for model M_1))/ (original c_d)	67.559/1.0035976 = 67.3168210	- -	- -
10. *Original* scaled ML difference test	((T_1 for model M_0) – (T_1 for model M_1))/ (original c_d)	65.473/1.0035976 = 65.2382987	65.477/1.0021854 = 65.3342186	65.557/0.8385757 = 78.1766035

Table 1 (continued)

Result	Statistic	Software Program		
		LISREL	**EQS**	**Mplus**
11. Model M_{10}	*df*	- -	19	19
	T_1	- -	108.453	108.587
	T_3	- -	91.8650	92.423
	c	- -	108.453/91.865 = 1.1805693	108.587/92.423 = 1.17489153
12. Term 2 for scaling factor of *new* scaled difference test	(*df* for model M_1) x (*c* for model M_{10})	- -	19 x 1.1805693 = 22.4308167	19 x 1.17489153 = 22.3229391
13. Scaling factor for the *new* scaled difference test (new c_d)	((*df* for model M_0) x (*c* for model M_0) – (*df* for model M_1) x (*c* for model M_{10}))/*m*	- -	(23.6495660 – 22.4308167)/1 = 1.2187493	(23.4130780 – 22.3229391)/1 = 1.0901389

Table 1 (continued)

		Software Program		
Result	**Statistic**	**LISREL**	**EQS**	**Mplus**
14. *New* scaled ML difference test	((T_1 for model M_0) – (T_1 for model M_1))/ (new c_d)	- -	65.477/ 1.2187493 = 53.7247488	65.557/ 1.0901389 = 60.1363734

Note. $N = 803$. M_1 = less restrictive baseline model. M_0 = more restrictive comparison model. T_1 = maximum-likelihood (ML) chi-square. T_2 = normal theory weighted least-squares (NTWLS) chi-square. T_3 = Satorra-Bentler scaled chi-square. c = scaling correction factor (Satorra & Bentler, 2001). m = (*df* for model M_0) – (*df* for model M_1). c_d = scaling correction factor for difference test. The above statistics are based on a comparison of factor loadings for items 5 and 11 from the Life Orientation Test (LOT; Scheier et al., 1994). For LISREL, the scaling correction factor (c) for each model is its NTWLS chi-square (T_2) value divided by its SB chi-square (T_3) value (Jöreskog et al., 1999). For EQS (Bentler, 1995) and Mplus (Muthén & Muthén, 2007), the scaling correction factor (c) for each model is its ML chi-square (T_1) value divided by its SB chi-square (T_3) value. For Mplus (Muthén & Muthén, 2007), the scaling correction factor (c) for each model is automatically reported to three decimals in the output when using the MLM estimator, and can be obtained to seven decimals using the RESULTS option in the SAVEDATA command. For LISREL, we report results for the original scaled difference test both when contrasting T_2 values and when contrasting T_1 values for models M_0 and M_1, although we recommend

that LISREL users contrast T_1 values (after recovering c for each model from T_2/T_3) to avoid obtaining an inadmissible negative value for the numerator of the scaled difference test. For EQS, we report results only when contrasting T_1 values for models M_0 and M_1. In computing the original scaled difference test, Mplus users should follow the same computational procedures as EQS users. In computing the new scaled difference test, we have omitted the results for LISREL 8.80 because the program produces an incorrect scaling correction factor for model M_{10} when the parameter estimates being contrasted are very different in magnitude. We have also omitted the value of T_2 for model M_{10} because the NTWLS chi-square value is only relevant for scaled difference testing via LISREL in recovering the value of c for model M_{10}.

Appendix A: Single-group LISREL, EQS, and Mplus syntax for estimating model M_{10}

LISREL syntax

```
!Two-factor CFA model for 8 scored LOT items: Model M10 for POOLED SAMPLE
!Testing the difference in the size of two factor loadings [for LOT items 5
!and 11 on Optimism Factor 1] using the new scaled difference test].
!Using final estimates from Model M0 as starting values with IT=0.
!Note that this model includes no equality constraint (EQ command).
DA NG=1 NI=8 NO=803 MA=CM
RA=LOT8.POOLED.PSF
ACM FI=LOT8.POOLED.ACM
SE
LOT1 LOT4 LOT5 LOT11
LOT3 LOT8 LOT9 LOT12 /
MO NX=8 NK=2 LX=FU,FR PH=SY,FR TD=DI,FR
PA LX
1 0
0 0
1 0
1 0
0 1
0 0
0 1
0 1
!The following parameter values for the LX, PH, and TD matrices have been
!manually copied and pasted from the final estimates for model M0, after
!replacing "- -" in the output for LX estimates with a value of "0.0" below.
MA LX
0.640     0.0
1.000     0.0
0.724     0.0
0.724     0.0
0.0       0.847
0.0       1.000
0.0       0.952
0.0       0.932
MA PH
0.740
-0.338    0.690
MA TD
0.889 0.242  0.541  0.622  0.540  0.299  0.303  0.399
LK
OPTIMSM PESSIMSM
OU SC ME=ML ND=3 IT=0
```

EQS syntax

```
/TITLE
Two-factor CFA model for 8 scored LOT items: Model M10 for POOLED SAMPLE
!Testing the difference in the size of two factor loadings [for LOT items 5
!and 11 on Optimism Factor 1] using the new scaled difference test].
!Using the final estimates from Model M0 as starting values with ITER=0.
!Note that this model includes no equality constraint.
/SPECIFICATIONS
DATA=LOT8.POOLED.ESS;
```

```
VARIABLES=8; CASES=803; GROUPS=1;
METHOD=ML,ROBUST; ANALYSIS=COVARIANCE; MATRIX=RAW;
/LABELS
V1=LOT1; V2=LOT4; V3=LOT5; V4=LOT11; V5=LOT3;
V6=LOT8; V7=LOT9; V8=LOT12;
!
! FOLLOWING LISTS ARE GENERATED FROM RETEST
!
/EQUATIONS
    V1  =  .641*F1  + 1.000 E1   ;
    V2  = 1.000 F1  + 1.000 E2   ;
    V3  =  .723*F1  + 1.000 E3   ;
    V4  =  .723*F1  + 1.000 E4   ;
    V5  =  .848*F2  + 1.000 E5   ;
    V6  = 1.000 F2  + 1.000 E6   ;
    V7  =  .952*F2  + 1.000 E7   ;
    V8  =  .933*F2  + 1.000 E8   ;
/VARIANCES
       F1=  .740* ;
       F2=  .690* ;
       E1=  .888* ;
       E2=  .242* ;
       E3=  .542* ;
       E4=  .621* ;
       E5=  .540* ;
       E6=  .299* ;
       E7=  .303* ;
       E8=  .399* ;
/COVARIANCES
       F2,F1  = -.338* ;
/PRINT
FIT=ALL;
TABLE=EQUATION;
/TECHNICAL
ITER=0;
/END
```

Mplus syntax

```
TITLE:    Two-factor CFA model for 8 scored LOT items: Model M10 for POOLED
          SAMPLE, testing the difference in the size of two factor loadings
          [for LOT items 5 & 11 on Optimism Factor 1] using the new scaled
          difference test -- using the final estimates from Model M0 as
          starting values with convergence=100000000. Note that this model
          includes no equality constraint.
DATA:     FILE=LOT8.POOLED.DAT;
VARIABLE: NAMES=LOT1 LOT4 LOT5 LOT11 LOT3 LOT8 LOT9 LOT12;
ANALYSIS: ESTIMATOR=MLM;
          convergence=100000000;
MODEL:    Optimism BY LOT4 LOT1;
          Pessimsm BY LOT8 LOT3 LOT9 LOT12;

!The following lines, taken directly from the output for model M0 when
!specifying the SVALUES option, fix the starting values of factor loadings.
!However, note that because model M10 should exclude the equality constraints
!added to model M0, we have deleted the equality constraint [i.e., the number
in parentheses (1) that was originally in the output for model M0], which
```

```
!indicated the loadings for LOT items 5 & 11 had been constrained to be equal
!in model M0.
     Optimism BY lot4@1;
     Optimism BY lot1*0.640;
     Optimism BY lot5*0.724;
     Optimism BY lot11*0.724;
     Pessimsm BY lot8@1;
     Pessimsm BY lot3*0.847;
     Pessimsm BY lot9*0.951;
     Pessimsm BY lot12*0.932;

!The following line, taken directly from the output for model M0 when
!specifying the SVALUES option, fixes the factor covariance:
     Optimism WITH Pessimsm*-0.338;

!The following lines, delimited by brackets and taken directly from the
!output for model M0 when specifying the SVALUES option, fix item
!intercepts:
     [ lot1*2.157 ];
     [ lot4*2.534 ];
     [ lot5*2.684 ];
     [ lot11*2.523 ];
     [ lot3*1.900 ];
     [ lot8*1.685 ];
     [ lot9*1.471 ];
     [ lot12*1.352 ];

!The following lines, taken directly from the output for model M0 when
!specifying the SVALUES option, fix item unique-error variances:
     lot1*0.888;
     lot4*0.241;
     lot5*0.540;
     lot11*0.621;
     lot3*0.539;
     lot8*0.299;
     lot9*0.303;
     lot12*0.399;

!The following lines, taken directly from the output for model M0 when
!specifying the SVALUES option, fix factor variances:
     Optimism*0.739;
     Pessimsm*0.689;

!Note that specifying the TECH5 option in the following OUTPUT command prints
!the iteration history, thereby allowing users to check to make sure they
!have set the convergence criterion high enough to halt iterations at zero.
OUTPUT: sampstat standardized tech1 tech5;
```

Appendix B: Applying the New Scaled Difference Test (Satorra & Bentler, 2010) in Multigroup CFA

Does optimism have the same meaning for men and women? As a multigroup example, we illustrate how to estimate model M_{10} in using the new scaled difference test to evaluate between-group factorial invariance. We use the same LOT-R data from the single-group example, first dividing respondents into separate groups of females (N = 647) and males (N = 156) for analysis via LISREL, EQS, and Mplus. Model M_1 freely estimates the loadings of the two-factor model of optimism for each gender, whereas model M_0 forces the factor loadings to be invariant with respect to gender. The difference in chi-square values between baseline model M_1 and nested model M_0 provides a test of the null hypothesis of gender invariance in factor loadings.

In conducting the new scaled difference test in a multigroup context, one sets up models M_1 and M_0 just as with the original scaled difference test. Model M_1 freely estimates the loadings of the baseline model for each group, whereas comparison model M_0 forces the factor loadings to be invariant with respect to gender. The difference in chi-square values between baseline model M_1 and nested model M_0 provides a test of the null hypothesis of gender invariance in factor loadings.

For LISREL, EQS, and Mplus, the multi-group CFA syntax for model M_{10} is identical to the multigroup syntax for model M_1, except for two modifications: (a) it includes a matrix of starting values for each group consisting of factor loadings, factor variances and covariance, and unique errors variances taken directly from the final parameter estimates in the output for model M_0 for each group; and (b) the number of iterations is frozen at zero.

Although the final estimates for each LISREL parameter matrix in model M_0 can be output to external files using the Output command in multigroup LISREL, the program stacks

matrix estimates for each group together, requiring users to split the parameter estimates from each group into separate external files. Also, as with single-group CFA, multigroup LISREL always exports the matrix of final estimates for unique error variances (Theta Delta) in a symmetric form, requiring users who specify the matrix of unique error variances as diagonal in model M_1 to respecify Theta Delta as a symmetric matrix with free diagonal elements and fixed subdiagonal elements for model M_{10}.

For these reasons, we recommend copying and pasting the final estimates from the output file for model M_0 for each group into the syntax file for model M_{10}, and then specifying these final estimates as starting values using MA commands for the Lambda-x, Phi, and Theta-Delta matrices in the multigroup CFA model. In addition, one must replaced the dashed lines (i.e. "- - ") reported for fixed values of zero in the Lambda-x matrix of the LISREL output for model M_0 with values of 0.0 in the matrix of starting values for Lambda-x in the syntax file for model M_{10}. As with single-group CFA, LISREL users can fix iterations at zero for multigroup model M_{10} by specifying IT=0 on the Output command line for each group. Below we present multigroup LISREL syntax for model M_{10}.

For multigroup EQS, users can obtain the starting values for model M_{10} by specifying a "retest" file (i.e., RETEST=*newfile*) in the PRINT section of the syntax file for model M_0, thereby storing for both groups the final parameter estimates of model M_0 in a separate outfile. EQS users can then copy and paste these final estimates for model M_0 directly from the retest file into the syntax for model M_{10} for each group. EQS users can fix iterations at zero for multigroup model M_{10} by specifying ITER=0 in the Technical section of the syntax file for each group. Below we also present multigroup EQS syntax for model M_{10}.

Mplus users can set up multigroup model M_{10} by using the option OUTPUT: SVALUES in the syntax file for model M_0 to generate syntax in the output file for model M_0 that copies each

group's final parameter estimates as starting values for M_{10}. However, note that because model M_{10} should exclude the invariance constraints added to model M_0, Mplus users must delete the numbers in parentheses included in the SVALUES output for model M_0, which indicate the equality-constrained parameters added to model M_0. Mplus users can freeze iterations at zero by specifying a sufficiently large convergence criterion (e.g., CONVERGENCE=100000000). Specifying the TECH5 option on the OUTPUT command prints the iteration history, thereby enabling users to inspect the Mplus output for model M_{10} to verify whether they have set the convergence criterion large enough to prevent iterations, or whether they must increase it to halt iterations.

Multigroup LISREL and EQS Syntax for Estimating Model M_{10}

LISREL Syntax

```
!Two-factor CFA model for 8 scored LOT items: Model M10 for FEMALES (GROUP 1)
!Testing the gender-invariance of factor loadings using the new scaled
!difference test. Using final estimates from Model M0 as starting values with
!IT=0. Note that this model includes no invariance constraints.
DA NG=2 NI=8 NO=647 MA=CM
CM FI=LOT8.FEMALE.cm
ACM FI=LOT8.FEMALE.acm
SE
LOT1 LOT4 LOT5 LOT11
LOT3 LOT8 LOT9 LOT12 /
MO NX=8 NK=2 LX=FU,FR PH=SY,FR TD=DI,FR
PA LX
1 0
0 0
1 0
1 0
0 1
0 0
0 1
0 1
!In each group, the following parameter values for the LX, PH, and TD
!matrices have been manually copied and pasted from the final estimates for
!model M0, after replacing "- -" in the output for LX estimates with a value
!of "0.0" below.
MA LX
0.640     0.0
1.000     0.0
0.932     0.0
0.537     0.0
0.0       0.849
0.0       1.000
```

```
0.0       0.944
0.0       0.935
MA PH
0.726
-0.317      0.662
MA TD
0.895  0.251  0.444  0.605  0.513  0.294  0.286  0.373
LK
OPT PESS
OU SC ME=ML ND=3 IT=0
!Two-factor CFA model for 8 scored LOT items: Model M10 for MALES (GROUP 2).
!Using final estimates from Model M0 as starting values with IT=0.
!Note that this model includes no invariance constraints.
DA NI=8 NO=156 MA=CM
CM FI=LOT8.MALE.cm RE
ACM FI=LOT8.MALE.acm RE
SE
LOT1 LOT4 LOT5 LOT11
LOT3 LOT8 LOT9 LOT12 /
MO NX=8 NK=2 LX=FU,FR PH=SY,FR TD=DI,FR
PA LX
1 0
0 0
1 0
1 0
0 1
0 0
0 1
0 1
MA LX
0.640     0.0
1.000     0.0
0.932     0.0
0.537     0.0
0.0       0.849
0.0       1.000
0.0       0.944
0.0       0.935
MA PH
0.718
-0.419      0.815
MA TD
0.854  0.285  0.484  0.702  0.653  0.320  0.389  0.477
LK
OPTIMSM PESSIMSM
OU SC ME=ML ND=3 IT=0
```

EQS Syntax

```
/TITLE
Two-factor CFA model for 8 scored LOT items: Model M10 for FEMALES (GROUP 1)
!Testing the gender-invariance of factor loadings using the new scaled
!difference test. Using the final estimates from Model M0 as starting values
!with ITER=0. Note that this model includes no invariance constraints.
/SPECIFICATIONS
DATA=LOT8.FEMALE.ESS;
VARIABLES=8; CASES=647; GROUPS=2;
```

```
METHOD=ML,ROBUST; ANALYSIS=COVARIANCE; MATRIX=RAW;
/LABELS
V1=LOT1; V2=LOT4; V3=LOT5; V4=LOT11; V5=LOT3;
V6=LOT8; V7=LOT9; V8=LOT12;
/EQUATIONS
!
! FOLLOWING LISTS ARE GENERATED FROM RETEST
!
/EQUATIONS   ! SECTION FOR GROUP  1
    V1  =  .640*F1  + 1.000 E1   ;
    V2  = 1.000 F1  + 1.000 E2   ;
    V3  =  .932*F1  + 1.000 E3   ;
    V4  =  .537*F1  + 1.000 E4   ;
    V5  =  .849*F2  + 1.000 E5   ;
    V6  = 1.000 F2  + 1.000 E6   ;
    V7  =  .944*F2  + 1.000 E7   ;
    V8  =  .935*F2  + 1.000 E8   ;
/VARIANCES   ! SECTION FOR GROUP  1
       F1=  .725* ;
       F2=  .662* ;
       E1=  .895* ;
       E2=  .251* ;
       E3=  .444* ;
       E4=  .605* ;
       E5=  .513* ;
       E6=  .294* ;
       E7=  .286* ;
       E8=  .373* ;
/COVARIANCES ! SECTION FOR GROUP  1
       F2,F1  = -.317* ;
/END
/TITLE
Two-Factor CFA model for 8 scored LOT items: Model M0 MALES (GROUP 2)
!Using final estimates from Model M0 as starting values with ITER=0.
!Note that this model includes no invariance constraints.
/SPECIFICATIONS
DATA=LOT8.MALE.ess;
VARIABLES=8; CASES=156;
METHOD=ML,ROBUST; ANALYSIS=COVARIANCE; MATRIX=RAW;
/LABELS
V1=LOT1; V2=LOT4; V3=LOT5; V4=LOT11; V5=LOT3;
V6=LOT8; V7=LOT9; V8=LOT12;
!
! FOLLOWING LISTS ARE GENERATED FROM RETEST
!
/EQUATIONS   ! SECTION FOR GROUP  2
    V1  =  .640*F1  + 1.000 E1   ;
    V2  = 1.000 F1  + 1.000 E2   ;
    V3  =  .932*F1  + 1.000 E3   ;
    V4  =  .537*F1  + 1.000 E4   ;
    V5  =  .849*F2  + 1.000 E5   ;
    V6  = 1.000 F2  + 1.000 E6   ;
    V7  =  .944*F2  + 1.000 E7   ;
    V8  =  .935*F2  + 1.000 E8   ;
/VARIANCES   ! SECTION FOR GROUP  2
       F1=  .717* ;
       F2=  .815* ;
```

```
       E1=  .854* ;
       E2=  .285* ;
       E3=  .484* ;
       E4=  .702* ;
       E5=  .653* ;
       E6=  .320* ;
       E7=  .389* ;
       E8=  .477* ;
/COVARIANCES ! SECTION FOR GROUP  2
       F2,F1  = -.419* ;
/TECHNICAL
ITER=0;
/PRINT
FIT=ALL;
TABLE=EQUATION;
/END
```

Mplus Syntax

```
TITLE:       Two-factor CFA model for 8 scored LOT items: Model M0 for
             FEMALES & MALES. Testing the gender-invariance of factor
             loadings using the new scaled difference test.
DATA:        FILE=LOT8.POOLEDwithGENDER.dat;
VARIABLE:    NAMES=LOT1 LOT4 LOT5 LOT11 LOT3 LOT8 LOT9 LOT12 GENDER;
             GROUPING=GENDER (0=FEMALE 1=MALE);
ANALYSIS:    ESTIMATOR=MLM;
             convergence=100000000;
MODEL:       optimism BY lot4;
             optimism BY lot1;
             optimism BY lot5;
             optimism BY lot11;
             pessimsm BY lot8;
             pessimsm BY lot3;
             pessimsm BY lot9;
             pessimsm BY lot12;

MODEL FEMALE:

!The following lines, taken directly from the output for model M0 when
!specifying the SVALUES option, specify the starting values of factor
!loadings for females. However, note that because model M10 should exclude
!the invariance constraints added to model M0, we have deleted the equality
!constraints [i.e., the numbers in parentheses that were originally in the
!output for model M0], which indicated that the non-fixed loadings for males
!and females had been constrained to be invariant in model M0.
     optimism BY lot4@1;
     optimism BY lot1*0.640;
     optimism BY lot5*0.932;
     optimism BY lot11*0.537;
     pessimsm BY lot8@1;
     pessimsm BY lot3*0.849;
     pessimsm BY lot9*0.945;
     pessimsm BY lot12*0.935;

!The following line, taken directly from the output for model M0 when
!specifying the SVALUES option, specifies the factor covariance for females:
     optimism WITH pessimsm*-0.316;
```

```
!The following lines, delimited by brackets and taken directly from the
!output for model M0 when specifying the SVALUES option, specify item
!intercepts for females:
     [ lot1*2.107 ];
     [ lot4*2.515 ];
     [ lot5*2.649 ];
     [ lot11*2.549 ];
     [ lot3*1.917 ];
     [ lot8*1.685 ];
     [ lot9*1.479 ];
     [ lot12*1.325 ];
     [ optimism@0 ];
     [ pessimsm@0 ];

!The following lines, taken directly from the output for model M0 when
!specifying the SVALUES option, specify item unique-error variances for
!females:
     lot1*0.894;
     lot4*0.250;
     lot5*0.443;
     lot11*0.604;
     lot3*0.513;
     lot8*0.294;
     lot9*0.285;
     lot12*0.372;

!The following lines, taken directly from the output for model M0 when
!specifying the SVALUES option, specify factor variances for females:
     optimism*0.725;
     pessimsm*0.661;

MODEL MALE:

!The following lines, taken directly from the output for model M0 when
!specifying the SVALUES option, specify the starting values of factor
!loadings for males. However, note that because model M10 should exclude the
!invariance constraint, we have deleted the equality constraints [i.e., the
!numbers in parentheses that were originally in the output for model M0],
!which indicated that the non-fixed loadings for males and females had been
!constrained to be invariant in model M0.
     optimism BY lot4@1;
     optimism BY lot1*0.640;
     optimism BY lot5*0.932;
     optimism BY lot11*0.537;
     pessimsm BY lot8@1;
     pessimsm BY lot3*0.849;
     pessimsm BY lot9*0.945;
     pessimsm BY lot12*0.935;

!The following line, taken directly from the output for model M0 when
!specifying the SVALUES option, specifies the factor covariance for males:
     optimism WITH pessimsm*-0.417;

!The following lines, delimited by brackets and taken directly from the
!output for model M0 when specifying the SVALUES option, specify item
!intercepts for males:
```

```
     [ lot1*2.365 ];
     [ lot4*2.609 ];
     [ lot5*2.827 ];
     [ lot11*2.417 ];
     [ lot3*1.833 ];
     [ lot8*1.686 ];
     [ lot9*1.436 ];
     [ lot12*1.468 ];
     [ optimism@0 ];
     [ pessimsm@0 ];

!The following lines, taken directly from the output for model M0 when
!specifying the SVALUES option, specify item unique-error variances for
!males:
     lot1*0.848;
     lot4*0.283;
     lot5*0.481;
     lot11*0.697;
     lot3*0.648;
     lot8*0.318;
     lot9*0.387;
     lot12*0.474;

!The following lines, taken directly from the output for model M0 when
!specifying the SVALUES option, specify factor variances for males:
     optimism*0.713;
     pessimsm*0.809;

!Note that specifying the TECH5 option in the following OUTPUT command prints
!the iteration history, thereby allowing users to check to make sure they
!have set the convergence criterion high enough to halt iterations at zero.
OUTPUT:      sampstat standardized tech1 tech5;
```