1

Measurement Error Adjustment for Single Indicator Variables

This primer discusses strategies for adjusting for measurement error in SEM when you have only a single indicator of a construct. I discuss two strategies, one based on fixed error variances and the other based on item parceling. The former method is described in depth in Savalei (2018); the latter method is described in Little, Rhemtulla, Gibson & Schoemann (2013).

THE FIXED ERROR VARIANCE APPROACH

The fixed error variance approach can be used when you have all single indicators in your model or a mixture of constructs with multiple and single indicators. The example I use to describe the method appears in Figure 1 and is an all-single indicator model. It is a semi-replication of the weight loss study reported on my website but it used a different population of overweight individuals, slightly different measures of the mediators, and a different time period over which weight loss was recorded (two weeks instead of one month). The treatment condition was represented by a variable called `TREAT` (0 = control group, 1 = treatment group). Weight loss was measured in pounds in the variable called `wloss`. It is the number of pounds lost over a two-week period. Positive numbers equal pounds lost and negative numbers equal pounds gained. There were two targeted program mediators, self-regulation (higher tendencies to self-regulate oneself leads to more weight loss) and self-efficacy for weight loss (higher confidence in one's ability to lose weight leads to more weight loss). Both self-regulation and self-efficacy were measured by the average response to multi-item inventories where each item was responded to on a 7 point disagree-agree scale (-3 = strongly disagree, -2 = moderately disagree, -1 = slightly disagree, 0 = neither, 1 = slightly agree, 2 = moderately agree, 3 = strongly agree). Total scores could thus range from -3 to +3 with higher scores indicating better self-regulation and higher self-efficacy, respectively. To keep the example simple, I omit covariates. I also assume you are reasonably familiar with Mplus and SEM.

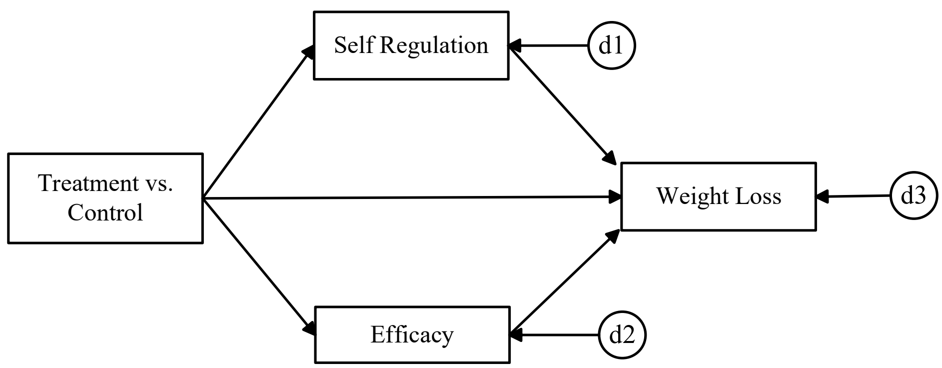

FIGURE 1. Single indicator weight loss example

The model in Figure 1 assumes all variables are measured without error. This is reasonable for participation in the treatment versus control conditions of an RET but not for the other variables. It is well known that measurement error can bias parameter estimates in linear regression, although the conditions under which it does so vary. A strength of multiple indicator SEM with interchangeable indicators for a construct is that it yields parameter estimates that take into account the measurement error theory of the investigator. However, we often encounter scenarios where, for practical reasons, we only have single indicators of constructs.

It is possible to recast the model in Figure 1 in latent variable terms. I have done so in Figure 2. Each observed measure that is thought to contain measurement error has a latent variable underlying it. The latent variable ostensibly represents the "true" construct underlying the measure of it. The observed measure serves as a reference indicator to define the metric of the latent variable, hence the path coefficient from the latent variable to the observed measure is fixed at 1.0. As noted, the model in Figure 1 assumes no measurement error in the continuous constructs and this is formalized in Figure 2 by setting the error variance of each indicator to zero. I do so by showing zeros next to the double headed arrows associated with the error variances of the indicators. With multiple indicators, we are able to estimate the error variance of the indicators, but we are unable to do so with single indicators because of under-identification; we have more unknowns than knowns in the model. Given this, researchers typically fix the error variances to zero rather than estimating them using Figure 1 and traditional Mplus syntax. I show the case where I assume there is no measurement error but make this explicit using latent variables via Figure 1. The models in the two figures will yield the same results:

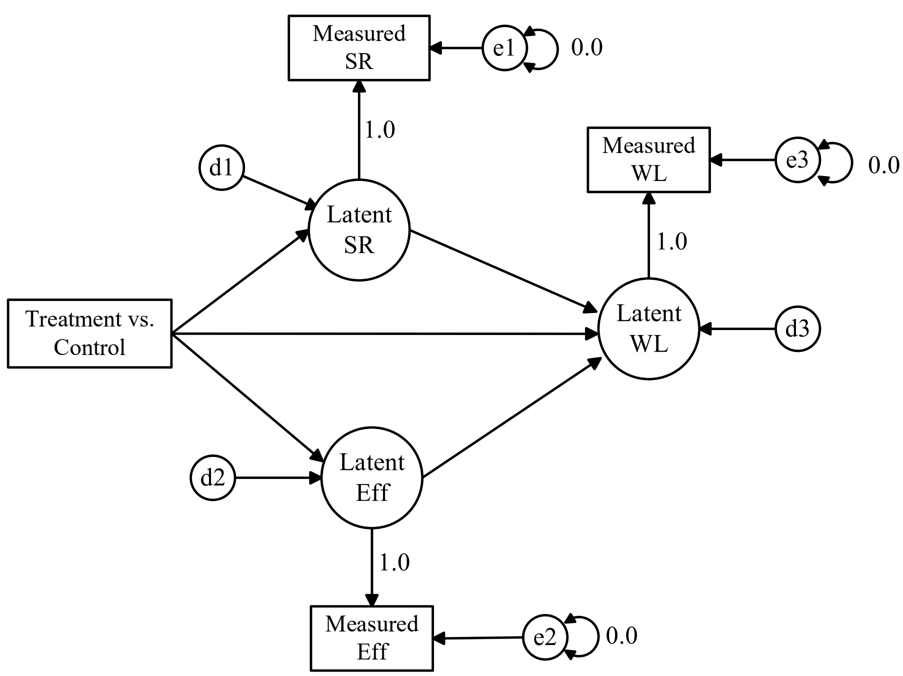

FIGURE 2. Single indicator weight loss example in latent variable terms

Table 1 presents the Mplus syntax for Figure 2. I number the lines for reference, but Mplus syntax excludes the numbers and the periods after them at the beginning of each line. The syntax could be more efficient but I sacrifice efficiency in the interest of pedagogy. I assume you have reviewed the basics of Mplus syntax on my website. Mplus is case insensitive.

Table 1: Mplus Syntax for Latent Variable with Single Indicator

```
1. TITLE: ANALYSIS OF SINGLE INDICATOR MODEL WITH LATENTS ;
2. DATA: FILE IS wloss.dat ;
3. VARIABLE:
4. NAMES ARE id wloss sr eff treat ;
5. USEVARIABLES ARE wloss sr eff treat ;
6. MISSING ARE ALL (-9999) ;
7. ANALYSIS:
8. ESTIMATOR=MLR ;
9. MODEL:
10. LSR BY sr@1 ;       !I use capital letter to denote latent variables
11. LEFF BY eff@1 ;     !when referring to latent variables in the syntax
12. LWLOSS BY wloss@1 ; !but this is optional;
13. sr@0 ;    !this refers to the error variance of the observed indicator sr
14. eff@0 ;   !this refers to the error variance of the observed indicator eff
15. wloss@0 ; !this refers to the error variance of the observed indicator wloss
16. [wloss@0]; [sr@0]; [eff@0] ; these are measurement intercepts fixed to zero
17.  [LSR] ; [LEFF] ; [LWLOSS] ; these are estimated latent intercepts
18. LSR ON treat ;  !This regresses LSR onto treat
19. LEFF ON treat ; !This regresses LSR onto treat
20. LWLOSS ON LSR LEFF treat; !This regresses LSW onto LSR, LEFF and treat
21. MODEL INDIRECT:
22. LWLOSS IND treat;
23. OUTPUT: SAMP STDYX MOD(All 4) RESIDUAL CINTERVAL TECH4 ;
```

Line 1 is the title line. Line 2 tells Mplus where to find the data file. Each line in the data file contains 5 values, space delimited, providing the scores for a given individual on the 5 input variables. Line 3 tells Mplus I am going to provide information about the variables that are in the data set. Line 4 provides the names I want to assign to the variables in the order they are encountered in the data file. There are 5 names because there are 5 variables. Line 5 specifies the subset of variables I want to use in the model. Line 6 tells Mplus that if it encounters the value -9999 for any of the variables, it should treat it as missing data. By default, Mplus uses full information maximum likelihood (FIML) for missing data for the endogenous variables. Line 7 tells Mplus I am going to provide information about the type of analysis I want. Line 8 specifies the estimator for the analysis to be robust maximum likelihood, to help deal with non-normality. Lines 9 tells Mplus I am going to provide information about the model per se. Line 10 tells Mplus that the latent variable I call `LSR` will be indicated by the observed variable called `sr`. I can name the latent variable anything, but I cannot have a name that exceeds 8 characters. The `BY` command is read as "is indicated by." The single indicator is listed after the `BY` and is followed by the `@` sign. The `@` sign is read as "fix the referenced parameter to a value of….", followed by the value you want to fix the parameter to. In this case, I fix the path from the `LSR` to `sr` to 1.0. I repeat the process in Lines 11 and 12 for `eff` and `wloss`. In Mplus, listing a variable by name refers to the variance of the variable or the error variance if the variable is endogenous. Line 13 fixes the error variance for `sr` to 0. Line 14 fixes the error variance of `eff` to 0. Line 15 fixes the error variance of `wloss` to 0. Line 16 contains each of the continuous indicators in brackets and refers to their measurement intercept when the indicator is regressed onto its latent variable. I constrain each of these to equal zero so that the mean of the indictor then becomes the mean of the latent variable, thereby furthering the equivalence between the observed measure and its latent variable. In Line 17, I estimate the intercepts of the three latent variables to further show the correspondence between the model in Figure 1 with the model in Figure 2. Line 18 tells Mplus to regress the latent variable `LSR` onto `treat`; line 19 tells Mplus to regress the latent variable `LEFF` onto `treat`; Line 20 tells Mplus to regress the latent variable `LWLOSS` onto `LSR`, `LEFF` and `treat`. Note that the core structural model focuses on the latent variables, not the observed indicators. Line 21 tells Mplus to conduct an analysis of mediation and Line 22 specifies the focus of the mediation analysis to be from the variable `treat` to the outcome `LWLOSS`. Finally, line 23 is the output line. I discuss the different options for the output line on the syntax tab of my webpage.

When I execute this syntax, I obtain the same results as for the model in Figure 1 with no latent variables. For reference, the traditional syntax for the Figure 1 model is shown in Table 2.[1] I now review the output for Table 1 and Figure 2.

Table 2: Mplus Syntax for Single Indicator Model

```
TITLE: ANALYSIS OF SINGLE INDICATOR MODEL ;
DATA: FILE IS wloss.dat ;
VARIABLE:
NAMES ARE id wloss sr eff treat ;
USEVARIABLES ARE wloss sr eff treat ;
MISSING ARE ALL (-9999) ;
ANALYSIS:
ESTIMATOR=MLR ;
MODEL:
sr ON treat ;
eff ON treat ;
wloss ON sr eff treat ;
MODEL INDIRECT:
wloss IND treat;
OUTPUT: SAMP STDYX MOD(All 4) RESIDUAL CINTERVAL TECH4 ;
```

The fit for the model was good. The chi square index of fit was 0.09 with 1 degree of freedom ($p < 0.77$), the CFI was 1.00, the RMSEA was <0.001 with a 90% confidence interval

[1] Technically, I do not need Lines 16 and 17 in Table 1 but if I want the output for the syntax in Table 1 to map onto the mean and intercept output for Table 2, I need to include them. In many applications, researchers are uninterested in the mean and intercept parameters in which case Lines 16 and 17 can be excluded.

of 0.000 to 0.089, the p value for close fit was 0.86 and the standardized RMR was 0.005. There were no meaningful modification indices larger than 4.

The output formatting for the model results for Figure 2 appears a bit differently than the standard output for Figure 1 because of the use of latent variables but, as noted, the numerical results are identical for model estimation of the corresponding parameters in the two models. Here is the portion of the output that focuses on the measurement model using the Table 1 output associated with Figure 2:

```
MODEL RESULTS

                                                    Two-Tailed
                    Estimate       S.E.  Est./S.E.    P-Value

 LSR      BY
    SR                 1.000      0.000    999.000    999.000

 LEFF     BY
    EFF                1.000      0.000    999.000    999.000

 LWLOSS   BY
    WLOSS              1.000      0.000    999.000    999.000
```

Note that, as planned, each of the factor loadings is fixed at 1.000 in the `Estimate` column. The standard errors in the second column are all zero, because the parameters were fixed not estimated. Mplus prints the value 999 when a statistic can't be computed, which is the case for the critical ratios and p values, again, because the parameter values they reference were fixed.

Here are the `Residual Variances` for the observed indicators of the latent variables:

```
Residual Variances

                                                    Two-Tailed
                    Estimate       S.E.  Est./S.E.    P-Value

    WLOSS              0.000      0.000    999.000    999.000
    SR                 0.000      0.000    999.000    999.000
    EFF                0.000      0.000    999.000    999.000
```

Each of the (measurement) error variances were a priori fixed at zero, as planned, so this output is as expected.

The path coefficients of interest in the model focus on the latent variables because it is at this level that the substantive causal relationships are specified. Here is the relevant output:

```
MODEL RESULTS

                                                    Two-Tailed
                    Estimate       S.E.  Est./S.E.    P-Value

 LSR      ON
    TREAT              0.986      0.199      4.968      0.000

 LEFF     ON
    TREAT              1.187      0.202      5.867      0.000

LWLOSS   ON
    LSR                0.874      0.141      6.183      0.000
    LEFF               1.056      0.151      6.990      0.000
    TREAT              0.468      0.678      0.691      0.490
```

The mean difference between the treatment and control conditions on the self-regulation mediator was 0.99 (critical ratio (CR) = 4.97, p < 0.001, margin of error (MOE) = ±0.40).[2] The mean difference between the treatment and control conditions on the self-efficacy mediator was 1.19 (CR = 5.87, p < 0.001, MOE = ±0.40). For the self-regulation mediator, for every one unit it increases, the mean weight loss is predicted to increase by 0.87 pounds, holding constant the other variables in the equation (CR = 6.18, p < 0.001, MOE = ±0.28). For the self-efficacy mediator, for every one unit it increases, the mean weight loss is predicted to increase by 1.056 pounds, holding constant the other variables in the equation (CR = 6.99, p < 0.001, MOE = ±0.30). Finally, the direct effect of the treatment condition on weight loss was statistically non-significant (CR = 0.69, ns). To summarize, the program significantly changed each mediator and each mediator was statistically significantly predictive of weight loss. Note that the results in this output are identical to the results I would obtain for the Table 2 syntax for Figure 1, driving home the fact that Figure 1 assumes no measurement error.

From, the mediation analysis output for the Table 1 syntax, I obtain the total effect of the treatment condition on weight loss. Here is the output:

```
TOTAL, TOTAL INDIRECT, SPECIFIC INDIRECT, AND DIRECT EFFECTS

                                                   Two-Tailed
                    Estimate       S.E.  Est./S.E. P-Value

Effects from TREAT to LWLOSS

  Total                2.584      0.659      3.920      0.000
  Total indirect       2.116      0.342      6.190      0.000
```

[2] For the margin of error, I use the informal calculation method of doubling the standard error.

The mean difference in weight loss over the two-week period between the treatment and control groups was 2.58 pounds (CR = 3.92, $p < 0.001$, MOE = ±1.32). This analysis also assumes no measurement error and the result will be the identical to what I would obtain from the syntax in Table 2.

The key to adjusting for measurement error for single indicators is to alter the Table 1 syntax so that instead of fixing the measurement error variances to zero, I fix them to values that reflect the presumed levels of unreliability of each measure. It turns out that in a single indicator latent variable model per Figure 2, the observed score variance of a measure will equal the latent variable ("true" score) variance plus the error variance:

var(X) = var(True) + var(Error)

It follows that the unreliability of a measure, in proportion terms, equals

Unreliability of X = var(Error)/var(X)

and through simple algebraic manipulation,

var(Error) = (Unreliability of X)*(var(X) [1]

Using Equation 1, I can create a table to calculate the amount of error variance in a measure that is associated with *what I believe* is the reliability of that measure. Table 3 presents such a table. I list in the first column the variance of each of the single indicator measures, taken from the Mplus output section labeled `UNIVARIATE SAMPLE STATISTICS` (or I can use the diagonal elements of the covariance matrix; or I can square the standard deviation from another software package). In the third column, I list what I think is the likely reliability of the measure, based on past research or pilot data. In column 4, I subtract from 1 the reliability index to obtain the unreliability index for the measure. Finally, in the last column, I multiply the Variance column by the Unreliability column to obtain the error variance associated with the measure.

Table 3: Error Variance Associated with Reliability

Measure	Variance	Reliability	Unreliability	Error Variance
sr	4.184	0.80	0.20	4.184*0.20 = 0.8368
eff	4.444	0.80	0.20	4.444*0.20 = 0.8888
wloss	45.123	0.90	0.10	45.123*.10 = 4.5123

Next, I change the syntax in Table 1 to fix the error variances for the measures not to zero but instead to the values in the last column of Table 3. The new syntax is shown in Table

4. It is identical to Table 1 except for the changes on Lines 13 to 15.

Table 4: Mplus Syntax for Latent Variable with Single Indicator

```
1. TITLE: ANALYSIS OF SINGLE INDICATOR MODEL WITH LATENTS ;
2. DATA: FILE IS wloss.dat ;
3. VARIABLE:
4. NAMES ARE id wloss sr eff treat ;
5. USEVARIABLES ARE wloss sr eff treat ;
6. MISSING ARE ALL (-9999) ;
7. ANALYSIS:
8. ESTIMATOR=MLR ;
9. MODEL:
10. LSR BY sr@1 ;
11. LEFF BY eff@1 ;
12. LWLOSS BY wloss@1 ;
13. sr@0.8368 ;
14. eff@0.8888 ;
15. wloss@4.5123 ;
16. [wloss@0]; [sr@0]; [eff@0] ;
17.  [LSR] ; [LEFF] ; [LWLOSS] ;
18. LSR ON treat ;
19. LEFF ON treat ;
20. LWLOSS ON LSR LEFF treat;
21. MODEL INDIRECT:
22. LWLOSS IND treat;
23. OUTPUT: SAMP STDYX MOD(All 4) RESIDUAL CINTERVAL TECH4 ;
```

When I execute the new syntax, I obtain new parameter estimates and significance tests that have been adjusted for the built-in levels of measurement error that I specified. I now highlight selected sections of the output to illustrate this.

The first section I examine is the standardized residuals for the targeted measures which shows the values of the unreliabilities for each one. I do this to ensure I did not make a mistake when calculating the error variance values. Here is the output:

```
STDYX Standardization

                                                Two-Tailed
                    Estimate       S.E.  Est./S.E.  P-Value

 Residual Variances
    WLOSS              0.100      0.007     14.011      0.000
    SR                 0.200      0.015     13.773      0.000
    EFF                0.200      0.014     14.610      0.000
```

If I calculated and entered values correctly, the `Estimates` should correspond to the unreliabilities in Table 3. This was indeed the case.

The global model fit indices were unchanged. This will often be the case, but not always, depending on the model and the type of constraints you impose.

Here are the key parameter estimates of interest:

```
MODEL RESULTS

                                                    Two-Tailed
                    Estimate       S.E.  Est./S.E.  P-Value

 LSR       ON
    TREAT              0.986      0.199      4.968      0.000

 LEFF      ON
    TREAT              1.187      0.202      5.867      0.000

 LWLOSS    ON
    LSR                1.109      0.185      6.012      0.000
    LEFF               1.349      0.196      6.898      0.000
    TREAT             -0.112      0.714     -0.157      0.876
```

The treatment versus control mean difference on self-regulation and self-efficacy are the same as in the analysis that assumed perfect reliability, but the estimated standard errors and critical ratios differ. The estimates for the effects of the mediators on weight loss are different. In the original analysis, the coefficient for self-regulation was 0.874 whereas here it is 1.109; for every one unit that self-regulation increases, the mean weight loss is predicted to increase by 1.109 pounds. In the original analysis, the coefficient for self-efficacy was 1.056. In the current analysis, it is 1.349; for every one unit that self-efficacy increases, the mean weight loss is predicted to increase by 1.349 pounds. The direct effect of the treatment condition on weight loss remains statistically non-significant. The standard errors and critical ratios for both mediators also differ when predicting weight loss.

From the mediation analysis, I obtain the estimated total effect of the treatment condition on weight loss. Here is the output:

```
TOTAL, TOTAL INDIRECT, SPECIFIC INDIRECT, AND DIRECT EFFECTS

                                                      Two-Tailed
                    Estimate       S.E.    Est./S.E.    P-Value

Effects from TREAT to LWLOSS

  Total                2.584      0.659      3.920      0.000
  Total indirect       2.696      0.439      6.139      0.000
```

The `Total indirect` effect in the prior analysis was 2.116; in the current analysis, it is 2.696.

In sum, the results from a model that adjusts for measurement error can differ from those of a model that assumes no measurement error. In this example, although some of the parameter estimates changed, the fundamental conclusions did not. The initial results were robust to the biasing effects of measurement error in terms of the conclusions I draw. In the next section, I delve into some of the intricacies of applying this fixed error variance approach.

ISSUES IN APPLYING THE FIXED ERROR VARIANCE APPROACH

Choosing Reliability Levels

An important issue when using the fixed error variance method is how to choose a reliability level to impose on a measure. One strategy is to base the reliability on the psychometric history of the measure with comparable populations and contexts. However, sometimes we use measures that lack such history. If one is working with a unidimensional multi-item composite, another strategy is to use a reliability value based on the composite reliability from item analyses in the data. Such estimates, however, are just that – estimates; they are subject to error and can vary from sample to sample, an issue I discuss in more depth see below. Savalei (2018) suggests an approach where one simply uses an *a priori* specified reliability for all measures, such as 0.80. The chosen reliability is based on an educated guess of the upper bound reliability for all targeted measures in the model. Although Savalei provides simulation evidence in support of the approach, it seems crude in its failure to recognize that measures vary in their reliability, sometimes considerably so.

Westfall and Yarkoni (2016) suggest a sensitivity approach in which the results of the SEM model are compared across a range of plausible reliability values. Although I embrace this suggestion in principle, my own experience is that implementing it can be challenging. For example, if my model has 4 error-prone single indicators and if I consider three levels of reliability for each (e.g., low = 0.70, medium =0.80, high = 0.90), there are 3^4 = 81 different reliability combinations. Nor is it clear the bases upon which one would choose to make conclusions across the 81 scenarios. A compromise is to select a prototypical "low reliability" profile, a "medium" reliability profile, and a "high reliability profile" based on one's prior knowledge of the measures and then test if the substantive conclusions are robust across the profiles. If the conclusions are robust, one moves forward with the conclusions with greater confidence. If not, one characterizes the necessary qualifications and moves forward with tentativeness.

I find that when I select reliability profiles that are low in reliability across all measures, I often encounter estimation problems, either in the form of non-convergence or wildly large standard errors. This can occur if one overcorrects for unreliability. For example, if the correlation between two variables is 0.75 and I assume both variables have a reliability of 0.70, the error corrected correlation between the variables turns out to be greater than 1.00, which is impossible, causing Mplus to give an error message. Sometimes the violations of plausible reliability are more subtle than this. For example, if the correlation between X and Z is 0.50 and the correlation between X and Y also is 0.50, then the correlation between Z and Y must be between -0.06 and 0.57. The use of erroneously low reliabilities might produce a Z-Y correlation outside this range, causing estimation difficulties in the SEM program. It usually is better to start with upper bound estimates of measure reliability as suggested by Savalei (2018) and then to explore robustness of conclusions as one introduces graduated reliability degradation. This approach can be used to identify the boundary conditions of conclusions relative to measure reliability.

In sum, there is no simple, straightforward rule for choosing reliability levels in the fixed error variance single indicator approach. However, using traditional SEM is tantamount to assuming there is no measurement error in any of the measures, which seems to bold a proposition. My own approach is to start with a traditional analysis. However, I then test the robustness of my conclusions as I progressively degrade measure reliability across plausible reliability scenarios, being careful to avoid over-correction.

Incorporating Uncertainty

Sometimes when we fix error variances to a given value, we are reasonably certain of the true population value of the error variance. If I calculate the error variance value based on a reliability estimate derived from sample data that has a very large N, then there likely is little sampling error associated with that estimate; I am reasonably certain of the error variance value I decide to specify even though it is from sample data. By contrast, if the estimate is based on a smaller N, the reliability estimate is subject to non-trivial sampling error. Now there is uncertainty about the correct value at which to fix the error variance. Oberski and Satorra (2013) argue that the standard errors for a parameter estimate corrected by uncertain reliability estimates should reflect the extent to which the reliability estimates are subject to sampling error. They suggest an inflation factor that can be incorporated into the estimation of parameter standard errors to reflect this dynamic. Importantly, the inflation factor can range from trivial to sizeable depending upon (a) the model structure and (b) the degree of sampling error in the reliability estimate. Oberski and Satorra's approach requires formal quantification of the sampling error associated with error variances which can be difficult in practice. However, their work is important because it underscores the need to appreciate the possibility that standard errors derived using the fixed error variance approach can be affected when relying on data-dependent methods to estimate values at which to fix the error variances. Their analysis also underscores the utility of sensitivity analyses using different reliability scenarios.

Bayesian Corrections

If you are using Bayes estimation, then the strategy of fixing measurement error variance for a single indicator to a non-zero value works but you cannot fix it to zero; this wreaks havoc with Bayesian estimation. If for whatever reason you want to fix it at zero (for example to use the Mplus `XWITH` command for a single indicator), you must instead fix the measurement error variance to a very small value close to zero. Asparouhov & Muthén (2020) recommend using a value that is between 1% and 3% of the total variance of the observed variable. Dor further discussion of this issue, see Asparouhov & Muthén (2020).

Addressing Systematic Measurement Error

The above discussion focuses on the case of random measurement error. For the case of systematic measurement error, once can use analogs of the fixed error variance approach. Suppose for example, that I wanted to test for effect sensitivity to the presence of correlated measurement errors in Figure 2. For example, I might hypothesize that social desirability tendencies impact both the self-regulation and self-efficacy measures of the mediators, which would create a correlation between e1 and e2 (assuming the variances of e1 and e2 have been fixed to non-zero values). It is possible to fix the covariance between fixed error variances to explore robustness of conclusions across correlated errors. The covariance between two errors equals

$$cov(e1,e2) = r_{e1e2}\ sd_{e1}\ sd_{e2}$$

where cov(e1,e2) is the covariance between e1 and e2, r_{e1e2} is the correlation between e1 and e2, sd_{e1} is the presumed standard deviation of e1, and sd_{e2} is the presumed standard deviation of e2. If the error variance of self-regulation is fixed at 0.8368 and the error variance of self-efficacy is fixed at 0.8888, then their respective standard deviations are the square roots of these values, 0.9147 and 0.9428. To fix the correlation between the errors at 0.20, this translates into a covariance of (0.20)(0.9147)(0.9428) = 0.172. I would then add a `WITH` statement to the model commands of Mplus that fixes the relevant covariance to 0.172, namely

```
sr WITH eff@0.172 ;
```

This will then adjust for the presumed level of correlated error. The same challenges described earlier apply to this strategy, but in more complex ways. The approach should be used with caution but it is, in principle, possible.

ITEM PARCELS

Another approach to adjust for measurement error when using a multi-item single indicator is to create item parcels. If I have a 30-item scale, I might divide it into three sets of 10 items and then form a composite score for each of these 10 item "parcels" of items. I then use the composite scores of the three parcels as separate indicators of the target latent construct and apply multiple-indicator SEM to address measurement error when estimating parameters. Although each parcel composite will be less reliable than the full 30-item composite, SEM takes this into account when calculating parameter estimates.

The literature surrounding the use of parceling is complex and filled with conflicting advice. In back-to-back articles in the same issue of *Psychological Methods*, Marsh et al. (2013) published an article titled "Why item parcels are (almost) never appropriate: Two wrongs do not make a right; camouflaging misspecification with item parcels in CFA models" followed immediately by an article by Little et al. (2013) titled "Why the items versus parcels controversy needn't be one." Interestingly, neither article mentioned the other. I found it instructive to read both articles, but I tend to resonate to the Little et al. summary of their viewpoint (p. 285):

> "*Parcels, per se, are not inaccurate, incorrect, or faulty. When thoughtfully composed, parcels provide efficient, reliable, and valid indicators of latent constructs. By considering the sources of variance of the items that go into parcels, including construct variance, specific variance, and measurement error, researchers can construct parcels with good measurement properties that can clarify the relations among latent variables. Parcels are not always appropriate and they are not always implemented correctly (Bandalos & Finney, 2007); we argue that these situations are no reason to remove this measurement tool from a researcher's arsenal of techniques.*"

As noted in Chapter 3 of my book, composites can be either unidimensional or multidimensional in character. I focus first on the case where the scale to be split into parcels is functionally unidimensional, an area where Little et al. and Marsh et al. tend to agree. Relevant to my discussion are psychometric distinctions between parallel indicators/items, tau equivalent indicators/items, and congeneric indicators/items. **Parallel indicators** have equal amounts of true-score variance and equal amounts of error variance. A factor analysis of them will reveal equal loadings and equal error variances. **Tau-equivalent indicators** have equal amounts of true score variance but can have different amounts of error variance. This also can be revealed in a factor analysis of the items. **Congeneric indicators** share true-score variance but the amount of such variance varies across indicators. Congeneric indicators also can have different error variances. The best strategy to form parcels differs depending on whether items of the broader test are parallel, tau-equivalent, or congeneric, as I elaborate shortly.

One issue when creating parcels is how many parcels to create. Little at al., (2013) argue for using a just-identified measurement space, which implies the use of three parcels. Marsh et al. (2013) prefer more indicators, hence more than three parcels. Both groups agree that creating two indicator/parcel models runs the risk of empirical under-identification and convergence difficulties, especially with smaller sample sizes. If the number of items is large, forming a larger number of parcels than three is viable. However, I often find that I am in situations were this is not possible because my multi-item scales have too few items.

Marsh et al. lean towards avoiding composites/parcels altogether and using all the single items as indicators of a single latent construct when modeling relationships between latent variables. Unfortunately, this is unrealistic in practice given sample sizes typical of RETs. If I have five latent constructs measured by composites of 10 items each and I measure them each at baseline, posttest, and follow-up, I must contend with a 150X150 covariance matrix. This is too large for most RETs.

Sometimes single items have relatively small amounts of true score variance and large amounts of random error. When aggregated into a composite, the true score variance accumulates and the positive and negative random errors cancel. As such, when there are more items in a parcel, the parcel tends to have higher factor loadings and less random measurement error, especially if the items approximate parallel indicator properties. Studies have shown that factor analytic and multiple-indicator SEM models often fare better when items/indicators have large loadings and small error variances (e.g., MacCallum et al, 1999). This favors not making the parcels too sparse in terms of the number of items they contain and questions Marsh's preference for using each item as a single indicator in its own right.

Different strategies have been suggested for segregating items into parcels. One popular but imperfect approach is that of random assignment of items to parcels with the constraint that parcels have roughly an equal number of items (Bandalos & Finney, 2001; Williams & O'Boyle, 2008). This strategy works best when there are many items that are parallel indicators with high loadings and do not have correlated errors, i.e., one can assume items are approximately interchangeable. Sterba and MacCallum (2010) have shown that structural model parameter estimates from one parcel allocation to the next can vary depending on how the random draw comes out. They recommend replicating results across different random draws to assure results are stable.

Instead of random assignment of items, a different approach involves **balancing** (Little et al., 2013). Balancing pairs the item with the highest loading with the item with the lowest loading into the first parcel; the next highest loading item and next lowest loading item are paired into the second parcel; the third highest and third lowest loading items are paired to form the third parcel. If there are three parcels and nine items, the process would be repeated for the next three item pairs but this time reversing the order of the parcel number in which the pairs are assigned. The general logic is that an item with a high loading reflects strong support for the construct so we match that with a weaker item. Across parcels, parallel or tau equivalence should then be approximated. This method has been found to work reasonably well for unidimensional scales (Rogers & Schmitt, 2004; Yang, Nay, & Hoyle, 2010).

Another issue to keep in mind for unidimensional scales is the need to assure that the content universe of items is evenly distributed within parcels. One wants to avoid one parcel having all items of one type while another parcel has all items of a different type. Landis et al. (2000) found that having researchers place items into parcels based on theory tended to result in more bias than using an empirical approach. Ideally, both empirical and theoretical information are used jointly to inform parceling (Little et al., 2013). Both Marsh et al. (2013) and Little et al. (2013) agree that parceling should not be done on a strictly *a priori* basis. Rather, researchers should engage in thorough item analyses and ensure they fully understand the psychometric properties of the items to decide an appropriate parceling strategy.

Strategies for parceling multi-dimensional scales are more controversial (Graham, Tatterson, & Widaman, 2000; Graham, 2004; Hall et al., 1999). One strategy that does not assume unidimensionality is **correlational parceling** in which items that correlate most strongly with one another are assigned to a parcel (Landis et al., 2000; Rogers & Schmitt, 2004). Another strategy is known as **facet-representative parceling** in which items that share facet-relevant content are assigned to the same parcel (Little et al., 2002; Kim & Hagtvet, 2003). Marsh et al. (2013) are skeptical about parceling strategies for multi-component constructs.

Several researchers have argued that parcels with SEM are almost always a better choice than using fully aggregated scales that ignore the presence of measurement error (Coffman & MacCallum, 2005), but the evidence for this argument is somewhat mixed if sample sizes are small (Ledgerwood & Shrout, 2011). I recommend you read the Marsh et al. (2013) and Little et al. (2013) articles for further elaboration of core issues in parceling. A nice summary of relevant issues for parceling and a careful analysis of the advantages and disadvantages of parceling is presented in Little, Rioux, Odejimi and Stickley (2022).

CONCLUDING COMMENTS

Measurement error can produce non-trivial bias when evaluating causal models (Cole & Preacher, 2014). This primer reviewed two methods for addressing it in the context of single indicator models, a fixed error variance approach and item parceling. Both have strengths and weaknesses.

REFERENCES

Asparouhov, T. & Muthén, B. (2020). Bayesian estimation of single and multilevel models with latent variable interactions, *Structural Equation Modeling*, 28, 314–328.

Bandalos, D, & Finney, S. (2001). Item parceling issues in structural equation modeling. In: Marcoulides, G.& Schumacker, R. (Eds.) *New developments and techniques in structural equation modeling*. p. 269-296. Erlbaum.

Coffman, D. & MacCallum, R. (2005). Using parcels to convert path analysis models into latent variable models. *Multivariate Behavioral Research*, 40, 235–259.

Cole, D. & Preacher, K. (2014). Manifest variable path analysis: Potentially serious and misleading consequences due to uncorrected measurement error. *Psychological Methods*, 19, 300-315.

Graham, J. (2004). Creating parcels for multi-dimensional constructs in structural equation modeling. Paper presented at the biennial meeting of the Society for Multivariate Analysis in the Behavioral Sciences, Jena, Germany.

Graham, J., Tatterson, J. & Widaman, K. (2000). Creating parcels for multidimensional constructs in structural equation modeling. Paper presented at the annual meeting of the Society of Multivariate Experimental Psychology, Saratoga Springs, New York.

Hall, R., Snell, A., & Foust, M. (1999). Item parceling strategies in SEM: Investigating the subtle effects of unmodeled secondary constructs. *Organizational Research Methods*, 2, 233–256.

Kim, S. & Hagtvet, K. (2003). The impact of misspecified item parceling on representing latent variables in covariance structure modeling: A simulation study. *Structural Equation Modeling*, 10, 101–127.

Landis, R., Beal, D., & Tesluk, P. (2000). A comparison of approaches to forming composite measures in structural equation models. *Organizational Research Methods*, 3, 186–207.

Ledgerwood, A. & Shrout, P. (2011). The trade-off between accuracy and precision in latent variable models of mediation processes. *Journal of Personality and Social Psychology*, 101, 1174–1188.

Little, T., Cunningham, W., Shahar, G. & Widaman, K. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling*, 9, 151–173.

Little, T., Rhemtulla, M., Gibson, K. & Schoemann, A. (2013). Why the items versus parcels controversy needn't be one. *Psychological Methods*, 18, 285–300.

Little, T., Rioux, C., Odejimi, O. & Stickley, Z, (2022). *Parceling in structural equation modeling.* Cambridge University Press

MacCallum, R. C., Widman, K. F., Zhang, S., and Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, 4, 84-89.

Marsh, H. W., Lüdtke, O., Nagengast, B., Morin, A. J. S., & von Davier, M. (2013). Why item parcels are (almost) never appropriate: Two wrongs do not make a right—Camouflaging misspecification with item parcels in CFA models. *Psychological Methods*, 18, 257–284.

Oberski, D. L., & Satorra, A. (2013). Measurement error models with uncertainty about the error variance. *Structural Equation Modeling*, 20, 409–428.

Rogers W. & Schmitt, N. (2004). Parameter recovery and model fit using multidimensional composites: A comparison of four empirical parceling algorithms. *Multivariate Behavioral Research*, 39, 379–412.

Savalei, V. (2018). A comparison of several approaches for controlling measurement error in small samples. *Psychological Methods*, 24, 352-370.

Sterba S. & MacCallum, R. (2010). Variability in parameter estimates and model fit across repeated allocations of items to parcels. *Multivariate Behavioral Research.*, 45, 322-358.

Westfall, J. & Yarkoni, T. (2016). Statistically controlling for confounding constructs is harder than you think. *PLoS ONE*, 11, e0152719.

Williams L., & O'Boyle, E. (2008). Measurement models for linking latent variables and indicators: A review of human resource management research using parcels. *Human Resource Management Review*, 18, 233-242.