

TESTING FOR GROUP DIFFERENCES IN ORDINAL MEASUREMENT MODELS WITH AND WITHOUT THE MPLUS INVARIANCE SHORTCUT

Measurement invariance exists when the parameters of a measurement model are statistically equivalent across two or more groups. Invariance suggests that the same scores on a scale reflect the same levels of the latent construct across groups. The measurement parameters of interest in invariance testing with ordinal data are the factor loadings and thresholds of indicators. Invariance in residual variances or scale factors (variances of the continuous latent distributions underlying categorical indicators) can also be tested but is not required in most social science research. Researchers typically hope to demonstrate their measures are invariant across groups. When loadings and thresholds are invariant across groups, scores on latent variables can be compared across the groups and the groups can be analyzed together in tests of hypotheses about relationships among latent variables. Below we describe possible steps in invariance testing with Mplus. It should be noted that there are differing views in the literature about the best ways to test invariance, including how to evaluate differences in fit across groups. The literature suggests that questions remain about the best methods for testing and evaluating the invariance of scales with ordinal variables (Sass, 2010). The number of technical simulation studies evaluating the effects on parameter estimates, standard errors, and fit indices of ordinal modeling strategies under different conditions continues to grow (Bauer, 2017; Lai, 2020; Liu & Sriutaisuk, 2021; Sass et al., 2014; Wu & Estabrook, 2016)). However, the studies have not resulted in any consistent, agreed upon recommendations for social scientists to follow. This document reports on the invariance model specifications available in Mplus 8.4 (Muthén & Muthén, 2019) and described in the Mplus User's Guide (Muthén, 1998-2017), The Mplus steps of testing are consistent with recommendation by numerous social science researchers (e.g., Brown, 2015; Dimitrov, 2010; Sass, 2011). Mplus invariance procedures rely on χ^2 difference tests, which is supported by a 2014 simulation study (Sass et al., 2014). An update to Mplus 7.11 provided a convenient shortcut for conducting a sequence of increasingly restrictive measurement invariance models (Muthén, 2013). Both the manual testing steps and the shortcut steps are included here along with syntax.

This handout presents:

- ❖ Background information on levels of invariance,
- ❖ Background information on how to obtain the proper analysis characteristics for ordinal data in Mplus,
- ❖ An illustration of the traditional steps in testing for measurement invariance with ordinal data in Mplus
- ❖ An illustration of how the Mplus invariance shortcut can simplify invariance testing (starting on page 13)

Three documents (Bowen, 2021a, 2021b; Bowen & Masa, 2015) listed in the Reference list provide additional information and guidance on invariance testing in Mplus.

Levels of Measurement Invariance for Ordinal Data

From Dimitrov (2010) we have the following definitions of four levels of invariance for ordinal data:

1. **Configural** Invariance

Groups have the same number of factors and indicators and the same pattern of loadings. Configural invariance should be established before researchers proceed to conducting measurement invariance tests.

Three levels of Measurement Invariance

2. **Metric** or Weak Invariance

In addition to configural invariance, the groups have invariant factor loadings.

3. **Scalar** or Strong Invariance

In addition to configural and metric invariance, the groups have invariant thresholds.

4. **Strict** or Uniqueness Invariance

In addition to configural, metric, and scalar invariance, the residual variances of items or scale factors are invariant. Strict invariance does NOT have to be demonstrated to claim measurement invariance.

Mplus Specification of Ordinal Measurement Models

For factor models (single or multiple group) based on ordinal indicator variables, a robust weighted least squares (WLS) estimator and polychoric correlation input matrix are recommended (Bollen, 1989). To obtain these analysis characteristics in Mplus, the CATEGORICAL ARE: statement is used in the VARIABLES section, and the WLSMV estimator is requested in the ANALYSIS: section of the input file. For example:

```
VARIABLES:
```

```
  NAMES ARE: . . . .;
```

```
  USEVARIABLES ARE  wb1 wb2 wb3 wb4 wb5;
```

```
  CATEGORICAL ARE  wb1 wb2 wb3 wb4 wb5;
```

```
  MISSING ARE ALL (99);
```

```
ANALYSIS:
```

```
  ESTIMATOR IS WLSMV;
```

It is possible to test for invariance across several groups at a time, but the testing can quickly become quite complicated. The example presented here includes just two groups. We also recommend testing one factor at a time for invariance. Our model is simple and manageable, but identifying noninvariant parameters in models with multiple factors and variables with many response options can become onerous. During invariance testing, most of our attention is on the comparison of fit of more and less constrained models. After establishing invariance, we must of course also evaluate fit and parameter estimates of the final scalar model. **Note: in all SEM testing, it is important to examine the output to ensure that Mplus estimated the model you think you specified. Mplus has defaults you may not be aware of. Parameter defaults are listed in the Mplus User's guide, but it is easier to just examine your output.**

Traditional Steps for Testing Measurement Invariance in Mplus

*(This section helps you understand the testing process, but you'll find **the shortcut method** on p.13 quicker and easier.)*

There is more than one way to conduct invariance testing. We present a sequence of steps based on the recommendations of many invariance experts and on our own extensive experience with invariance testing.

STEP 1: Test the hypothesized CFA model separately in each group.

Invariance scholars recommend obtaining the best factor model for each group separately before beginning invariance testing (Dimitrov, 2010). To run separate analyses for groups represented by different values on a variable in the data set, use the following syntax in the VARIABLE section of the code, substituting your grouping variable's name and the value in the dataset that represents the group you want to examine:

```
USEOBSERVATION ARE (raceth EQ 3);          (for analyses without a cluster variable)
OR
SUBPOPULATION IS (raceth EQ 3); (for analyses with cluster variable and Type=Complex)
```

If data for your groups are in separate data files, specify the data file for one group at a time in the usual way:

```
DATA: FILE IS "C:\MyMplus\datafilewhite.dat";
```

We identify the best factor structure for each group in a single group analysis by running the hypothesized model and examining its fit and possibly its modification indices. To obtain modification indices add a line of code in the input file:

```
OUTPUT: MODINDICES (3.84);
```

See also the document *Requesting and Using Modification Indices in Mplus* (Bowen, 2021c). Group-specific models may include different correlated errors, but the pattern of factors and loadings should be essentially the same for the two groups. We hope for a good-fitting model for each group, but even if fit is marginal for a group, it is okay to proceed to invariance testing across groups—as long as the factor structures of the models are essentially the same. Note that Modification Indices are not available during the invariance tests below.

STEP 2: Request a multiple group model in Mplus (basics)

Multiple group models can be run from multiple datasets or based on values on a grouping variable in one dataset. To specify a multiple group model based on values on a variable in the dataset, use code like the following.

```
GROUPING IS raceth (1 = AA 3 = WH );
```

The "GROUPING IS" code lets Mplus know you want to do a multiple group analysis and indicates which values on the grouping variable represent the groups of interest. The code

goes in the VARIABLE: section of the input file. Note: The grouping variable is not put in the USEVARIABLE statement because it has a “special function.”

To specify that group data are in two separate files, we would include two file statements:

```
DATA: FILE (AA) IS AA.dat;  
      FILE (WH) IS white.dat;
```

The file statement for each group includes in parentheses the name you want Mplus to use when reporting results for each group.

If the configural models for the two groups are exactly the same (based on the single group tests), we simply enter one model statement in the usual way. The model statement below specifies that a factor called WELLBEING is measured by five observed indicators in both groups. Because no code appears under the MODEL WH: heading, we are testing the same model in both groups.

```
VARIABLES:  
  NAMES ARE: . . . .  
  USEVARIABLES ARE  wb1 wb2 wb3 wb4 wb5;  
  CATEGORICAL ARE  wb1 wb2 wb3 wb4 wb5;  
  MISSING ARE ALL (99);  
  GROUPING IS RACETH (1 = AA 3 = WH );  
ANALYSIS: ESTIMATOR IS WLSMV;  
MODEL:  
  WELLBEING by wb1 wb2 wb3 wb4 wb5;  
MODEL WH:
```

We specify differences in the two groups’ models by using a separate MODEL statement for the second group. If a correlated error was found between wb3 and wb4 in the WH group but not in the AA group, for example, it could be specified as follows:

```
MODEL:  
  WELLBEING BY wb1 wb2 wb3 wb4 wb5;  
MODEL WH:  
  wb3 WITH wb4;
```

Group specific model statements indicate parameters that *differ* from the parameters given in the general model statement. The parameters specified in this section of the code will be freely estimated in both groups; the code removes the default equality constraint that would pertain if only the general model statement were used. If the correlated error appeared in the AA group instead of the WH group, the WH model statement could be modified to indicate the correlated error does NOT apply to the WH group:.

```
MODEL:
    WELLBEING BY wb1 wb2 wb3 wb4 wb5;
    wb3 WITH wb4;
MODEL WH:
    wb3 WITH wb4@0;
```

To summarize: the general model statement refers to the first group (here AA) AND to subsequent groups EXCEPT for group specific statements.

STEP 3. Specify a baseline multiple group model with free loadings and free thresholds (a configural model).

By Mplus default, the code in Step 2 specifies a **scalar** model in which loadings and thresholds are constrained to be equal across the groups. Scalar models have more constraints than configural and metric models. To conduct a χ^2 difference test the traditional way in Mplus, the **less** constrained model has to be run first. Therefore, to create the baseline (least constrained model), we have to override the default constrained specification with code that frees a lot of loadings and thresholds.

For simplicity, we will look at a model without any correlated errors. The baseline model for invariance testing is a model with loadings and thresholds free to vary across the two groups—the configural model. However, for identification purposes, certain loadings and thresholds must be fixed or constrained to equality across groups. Specifically, at least one loading per factor should be fixed at 1, making it the “referent” or “marker” indicator. By default the first item listed in the BY statement is made the referent in Mplus, but the default can be overridden by making another indicator the referent when you want to test the equivalence of the first loading. (More on this process later.)

At least 2 thresholds have to be constrained to be equal across groups **for the referent indicator**. For other indicators, just one threshold must be constrained for identification purposes. By default Mplus constrains the first threshold for each indicator, but it does not matter which thresholds are constrained, as long as the minimum number is constrained for each indicator. The variables in the example have 3 response options and therefore 2 thresholds. We cannot test the thresholds for the referent indicator, currently wb1, because both have to be constrained for identification purposes. Later, however, we can make another variable the referent in order to test wb1’s thresholds.

The syntax below specifies a baseline model in which freeable loadings and thresholds are freed from the default equality constraints.

```
MODEL:
    WELLBEING BY wb1 wb2 wb3 wb4 wb5;
MODEL WH:
    WELLBEING BY wb2 wb3 wb4 wb5;
    [wb2$2];
    [wb3$2];
    [wb4$2];
    [wb5$2];
```

The MODEL: statement specifies the factor model that applies to both groups. By default the statement on its own imposes equality constraints on all loadings and thresholds. The MODEL WH: statements indicate how the model differs across groups. By repeating the model statement with four loadings, we instruct Mplus to keep wb1 fixed at 1 in the WH group (by not listing the variable) and request that the subsequent four factor loadings be estimated freely in each group (overriding the default equality constraints for loadings).

Important Note: If wb1 were included in the MODEL WH: statement, the model would not be identified. Listing wb1 would override the default fixing of the first indicator's loading to 1 in the WH group, leaving the group's factor model unidentified.

One threshold for each of the other indicators is also named in the MODEL WH: section. Thresholds are freed by bracketing their names under MODEL WH:. The threshold names are assigned by Mplus and should not be changed. Listing the term [wb2\$2]; frees the second threshold for item wb2. No thresholds are freed for wb1 because both have to be constrained to equality for identification. There are four other thresholds in the model—wb2\$1, wb3\$1, and wb4\$1, and wb5\$1. They cannot be tested for invariance at the same time as their \$2 counterparts, but the \$1s should also be tested if noninvariance is found in the initial test of the scalar model. Replacing \$2 with \$1 in a threshold name will free its \$1 threshold for testing. Mplus will then automatically constrain \$2 for identification.

Another additional line of code is required to conduct a difference test of the baseline model's fit relative to subsequent more constrained models. A small text file with information about the baseline model must be created and saved by Mplus. To request the creation of the file in our example, we use the code below.

```
SAVEDATA: DIFFTEST=configural.dat;
```

We have invented a name for the file that will help us remember it refers to the baseline model with free loadings and thresholds: "configural.dat." Now we can run this less restrictive configural model. The goal of the run is just to get that configural.dat file saved for the next step. (See also the handout " χ^2 Difference Testing in Mplus with WLSMV Estimation.")

STEP 4: Specify a multiple group model with constrained loadings and free thresholds (a metric model).

After Step 3, we run a more constrained model—in this case the metric model—to see if fit deteriorates significantly with loading constraints. In the model statement below we have turned off the statement freeing loadings for items wb2, wb3, wb4, and wb5 under MODEL WH:. We have kept the freeable thresholds free as is appropriate in a metric model. The loading for wb1 remains fixed at 1 in both groups, so it cannot be freed. Also in this step, we move the DIFFTEST statement to the ANALYSIS section of the input file. We delete the file name or turn off the previous SAVEDATA: command so we don't overwrite the original file (very important!).

ANALYSIS: **DIFFTEST = configural.dat;**

MODEL: WELLBEING BY wb1 wb2 wb3 wb4 wb5;

MODEL WH:

! WELLBEING BY wb2 wb3 wb4 wb5; “!” turns off the line of code, reinstating the default equality constraints for loadings.
 [wb2\$2];
 [wb3\$2];
 [wb4\$2];
 [wb5\$2];

SAVEDATA:

We delete the difftest instructions here.

The output from the model will include χ^2 difference test results for the comparison of the metric model and configural model. Difftest results with p values < .05 indicate significant deterioration in fit as measured by χ^2 . With the output presented below we would conclude at least one loading is noninvariant because $p < .05$. Note the change in degrees of freedom corresponds to the number of parameters we just constrained relative to the configural model—four loadings.

Mplus Difftest Output: Metric vs. Configural Model

Chi-Square Test for Difference Testing

Value	20.422
Degrees of Freedom	4
P-Value	0.0004

This is the change in χ^2 .

This is the change in df.

This is the p value of the change.

If loading invariance had been found here in Step 4, we could have proceeded to Step 6. Instead we proceed to Step 5.

STEP 5: Explore partial metric invariance. If noninvariance among loadings was revealed in

Step 4, specify a sequence of models with one loading constrained at a time.

We will begin testing one loading at a time to see if we can isolate the cause(s) of the noninvariance revealed in step 4. Using modification indices might speed up the process of finding noninvariant loadings; they would indicate which parameters **not** to constrain. The search for sources of noninvariance is based on the notion of “**partial invariance**” (Byrne et al., 1989). Partial invariance exists if only a small number of parameter estimates are noninvariant. Dimitrov (2010) suggests that if fewer than 20% of the loadings (in this case) are noninvariant, we can model the partial metric invariance and proceed to testing thresholds. In a model with only 5 loadings, one noninvariant loading could be tolerated; any more would put us above the 20% cutoff. So we are hoping to find no more than one noninvariant loading.

We will compare a model with all freeable loadings free (the configural model) to a model with one loading constrained at a time. With the traditional method in Mplus, we must run the less restrictive model first; the configural model here. With one loading constrained (aside from wb1), the second model is “nested” within the first model. A nested model has a subset of the freely estimated parameters of the first model. Nested models have fewer freely estimated parameters than the comparison model and, therefore, more degrees of freedom. Nested models are a requirement for chi square difference testing. If in any

multiple group testing you are unsure which model is nested in the other, run both models and compare the Number of Free Parameters given in the output. The model with more free parameters has fewer constraints. The other model is nested within it (as long as the constraints it adds are for parameters that were previously free).

We will compare the configural model to a model in which one loading is constrained to have the same value in both groups. Wb1 is currently fixed at 1, so we start by constraining the loading for wb2. We already have the difftest text file saved for the configural model (configural.dat), so we name it under the ANALYSIS command.

ANALYSIS:

DIFFTEST = configural.dat;

MODEL: WELLBEING BY wb1 wb2 wb3 wb4 wb5;

MODEL WH:

WELLBEING BY wb3 wb4 wb5;

[wb2\$2];

[wb3\$2];

[wb4\$2];

[wb5\$2];

SAVEDATA:

By not listing wb2, its default equality constraint is reinstated; all freeable thresholds are free.

We delete the DIFFTEST data file code to avoid overwriting the configural.dat file.

Results indicate that with wb2 constrained, fit deteriorates relative to the baseline comparison model. The conclusion is that the loading for wb2 is noninvariant.

**Mplus Difftest Output:
wb2 constrained**

Chi-Square Test for Difference Testing

Value 17.240

Degrees of Freedom 1

P-Value 0.0000

χ^2 increased with the additional constraint.
df increased by 1 when we constrained wb2.
 $p < .05$ so the constraint significantly hurt fit.

It is important to be systematic and thorough when doing invariance tests. We recommend testing one parameter at a time, writing down the change in χ^2 and its p value, then returning the parameter to its previous state (fixed or constrained) in the baseline before testing the next parameter. From the above Difftest Output we would note:

Constrain only wb2: chi sq diff = 17.2 p = .0000

Returning to our example: we test the effects of constraining just wb3, then just wb4, and then just wb5. We find:

Constrain only wb3: chi sq diff = 2.8 p = .0942

Constrain only wb4: chi sq diff = 6.7 p = .0098

Constrain only wb5: chi sq diff = 4.7 p = .0309

To test the invariance of the loading for wb1, we need to make a different indicator the referent. It is best to choose the most invariant indicator (i.e., the one associated with the smallest change in χ^2 and largest p value). Given our notes above, wb3 is the best choice for the referent. So we respecify the measurement of WELLBEING in **both groups** as follows:

MODEL: WELLBEING BY **wb3 wb1** wb2 wb4 wb5; Wb3 is now the referent.

MODEL WH:

WELLBEING BY wb2 wb4 wb5; We don't list wb3; its loading is now fixed to 1.
 We don't list wb1 because we are testing the effect
 of constraining its loading across groups.
 [wb2\$2];
 [wb1\$2];
 [wb4\$2]; The thresholds for wb3 remain constrained.
 [wb5\$2]; Thresholds for the other indicators are free.

DiffTest results indicate the loading for wb1 is invariant. The results so far indicate we have two invariant loadings (wb1, wb3) and three noninvariant loadings (wb2, wb4, wb5).

Constrain only wb1:	chi sq diff = 2.8	p = .0942
Constrain only wb2:	chi sq diff = 17.2	p = .0000
Constrain only wb3:	chi sq diff = 2.8	p = .0942
Constrain only wb4:	chi sq diff = 6.7	p = .0098
Constrain only wb5:	chi sq diff = 4.7	p = .0309

We recommend testing one loading (or threshold) at a time and recording its effect on χ^2 because the order in which items are tested can affect invariance results. For example, if we kept wb2's loading free after finding that constraining it led to a significant change in χ^2 , the result of a subsequent test of wb4 might be affected. There are times when freeing or constraining combinations of parameters may change the conclusions made based on the individual tests. That is, a loading previously found to be noninvariant might be found to be invariant if constrained and tested after one or two other more noninvariant loadings are freed. In these cases, your notes on χ^2 differences and their p values will be valuable. To find the combination of the fewest noninvariant parameters possible, free the two parameters determined to be most noninvariant based on diffTest results (those associated with the largest changes and most significant changes in χ^2). Run the model. If the diffTest is still significant, add the next most noninvariant parameter.

The choice of referent indicator can also affect invariance testing results. By testing each indicator separately, we can identify the most invariant loading (at least when the default referent is used). If we decide to change the default referent indicator after testing its invariance, the steps described above for retesting combinations of noninvariant loadings are even more important. The original referent may have affected results, but now we can defend our choice of referent. Note: In truth, we could repeat all individual tests with each indicator in turn serving as a referent (Cheung & Rensvold, 1999). It does not seem to be common practice and would be quite laborious. The steps above should be sufficient to get accurate and defensible results, especially with the step of testing combinations of noninvariant loadings after the best referent is put in place.

With only five loadings in our model, three noninvariant loadings represent 60% of the parameters, so we would have to conclude the construct WELLBEING is measured differently and means something different in the two groups. Out of curiosity, however, we will determine if freeing the two most noninvariant indicators, wb2 and wb4, helps change

the status of wb5. We free both wb2 and wb4, and conduct a difftest against the configural model. According to the output below, even with both wb2 and wb4 freed, the metric model remains significantly different from the configural model. Wb5 would also have to be freed to get that p value above .05

**Mplus Difftest Output: Constrained
wb2 & wb4 vs metric**

Chi-Square Test for Difference Testing

Value	6.4190
Degrees of Freedom	2
P-Value	0.0404

If metric noninvariance is found, as in our example, no more invariance testing is conducted; we must conclude the well-being measure is noninvariant across the two groups. If, in contrast, loading invariance (or partial invariance by the 20% rule) were established for the metric model, we would proceed to Step 6. To illustrate the process of testing thresholds, we will proceed to Step 6 as if we found full or partial metric invariance.

STEP 6: If full or partial invariance is found among loadings (metric invariance), test for invariance in thresholds (scalar invariance).

We test thresholds for invariance using the general procedures described above for loadings. The baseline comparison model for the threshold tests would be the final metric model from Step 4 (invariant) or Step 5 (partially invariant). We would have to run the final metric model with the SAVEDATA: request to obtain the appropriate text file for threshold difftests. The new file, which could be called for example, **metricmodel.dat**, would be referred in the ANALYSIS: section of the subsequent thresholds difftests.

With the metric model specified as appropriate and its difftest information saved, we constrain all thresholds to equality across groups and examine the effect on χ^2 .

ANALYSIS: DIFFTEST is metricmodel.dat;

MODEL: WELLBEING BY wb3 wb1 wb2 wb4 wb5;

MODEL WH:

!WELLBEING BY ;

We pretend we have metric invariance.

![wb2\$2];

We turn off the code for freeing thresholds.

![wb3\$2];

![wb4\$2];

![wb5\$2];

SAVEDATA:

The code freeing the four freeable \$2 thresholds is turned off, reinstating the default cross-group constraints of a scalar model. The two thresholds for wb1 and the \$1 thresholds for wb2 through wb5 are by default already constrained to be equal.

The results below from the model indicate the thresholds are invariant. Even with all thresholds constrained to equality across the two groups, fit does not deteriorate significantly relative to the metric model. After demonstrating scalar invariance (which by

definition includes our “pretend” metric invariance), we could conclude the construct is measured equivalently across the two groups.

Mplus DiffTest Output: Constrained Thresholds vs. metric

Chi-Square Test for Difference Testing

Value	4.292
Degrees of Freedom	4
P-Value	0.3679

The p value of the change in χ^2 is non-significant.

For demonstration purposes, however, we review what to do if the fully constrained scalar model evidenced noninvariance. Step 7 describes testing for partial invariance among thresholds.

STEP 7: Explore partial scalar invariance. If noninvariance among thresholds is revealed in Step 6, specify a sequence of models with one threshold constrained at a time.

If fit deteriorated significantly with all thresholds constrained, we would proceed to constrain one threshold at a time, as we did in Step 5 with loadings, to identify the source(s) of noninvariance. We want to compare χ^2 for a model with one threshold constraint to χ^2 for the metric model (in which all freeable thresholds were free). With loadings we tested one loading constraint against the configural model plus the default metric constraints (one loading fixed at 1 per factor). Here we test one threshold constraint against the final metric model plus the default scalar constraints (two thresholds constrained for the referent, one for each other indicator). We test a threshold’s invariance by **not freeing it** in the MODEL WH: section (i.e., keeping its default constrained status).

ANALYSIS: DIFFTEST is metricmod.dat;

MODEL:

WELLBEING BY wb1 wb2 wb3 wb4 wb5;

MODEL WH:

!WELLBEING BY ; We are treating the loadings as invariant.

! [wb2\$2]; Turning off this code reconstrains wb2\$2 for invariance testing.

[wb2\$1]; We free wb2\$1; it is freeable when wb2\$2 is constrained.

[wb3\$2];

[wb4\$2]; All other freeable thresholds are free.

[wb5\$2];

The above code specifies a model with more constraints (6) than freed (4) thresholds because the wb variables have only two thresholds. Specifically:

wb2\$2 is constrained; we turned off code freeing it.

wb1\$1	wb1\$2	} are constrained by default for identification	wb2\$1	} are free because we freed them.
wb3\$1			wb3\$2	
wb4\$1			wb4\$2	
wb5\$1			wb5\$2	

Because wb2\$2 is now constrained wb2\$1 is freeable, so we free it.

Because of required threshold constraints, we are testing the effect of one constraint at a time in the freest model possible, which is a model with multiple threshold constraints. We record the effect of constraining wb2\$2 in the context of all the default threshold constraints. The results indicate the threshold is invariant. We could proceed to test the effects of constraining the \$2 thresholds for wb3 through wb5 the same way.

```

Constrain only wb2$2:  chi sq diff = 2.8   p = .41
Constrain only wb3$2:  chi sq diff = ...   p = ...
Constrain only wb4$2:  chi sq diff = ...   p = ...
Constrain only wb5$2:  chi sq diff = ...   p = ...

```

Then we could test the \$1 thresholds in turn in the same way; see code below. Note we don't necessarily have to list all the \$2s or \$1s together. We could have tested wb2\$2, then wb2\$1, for example. The main idea is that we free as much of the threshold model as possible (one free threshold for each of our variables—either one will do) in order to test the effects on fit of constraining one threshold at a time. (With more than two thresholds per indicator, there would be more freeable thresholds.)

```

MODEL:
    WELLBEING BY wb1 wb2 wb3 wb4 wb5;
MODEL WH:
    !WELLBEING BY ;
    ! [wb2$1];      Constrain wb2$1 by not freeing it.
    [wb2$2];        Free wb2$2 because it is freeable with $1 constrained.
    [wb3$1];
    [wb4$1];        Free one threshold for the other indicators ($1s or $2).
    [wb5$1];

```

To test the thresholds of wb1, we would have to change the referent indicator in both groups as we did above to test wb1's factor loading. We identified wb3 as a good referent above, so we specify the model as follows:

```

ANALYSIS: DIFFTEST is metricmod.dat;
MODEL:
    WELLBEING BY wb3 wb1 wb2 wb4 wb5;      Wb3 is now the referent.
MODEL WH:
    !WELLBEING BY ;
    ! [wb1$1];      We can test the thresholds for wb1 now. We constrain
    [wb1$2];        wb1$1 and free wb1$2.
    [wb2$2];        No threshold for wb3 is listed because it is now the
    [wb4$2];        referent.
    [wb5$2];        One threshold is freed for each other indicator.

```

If invariance or partial invariance of loadings **and** thresholds is established according to the 20%, we have achieved scalar invariance—our goal. We would then proceed to examine the fit and parameter estimates of the final scalar model.

Although much of the traditional partial invariance testing procedures described in Steps 5 and 7 are also necessary with the Mplus Invariance Shortcut, many aspects of invariance testing are simplified with the shortcut.

The Mplus Invariance Shortcut

Starting with Version 7.11, Mplus has a shortcut (Muthén, 2013) that allows us to simultaneously run and compare χ^2 s for configural, metric, and scalar measurement models of two or more groups. We specify the grouping variable or two data files as usual and add the following line to the ANALYSIS section of the input file:

```
MODEL IS CONFIGURAL METRIC SCALAR;
```

The shortcut combines Steps 2, 3, 4, and 6 from the traditional approach into one analysis without us having to write code to specify the configural model with free loadings and thresholds or the metric model with its equality constraints on factor loadings but not thresholds. We also don't have to use the SAVEDATA: and ANALYSIS: difftest code. The model difftests happen automatically with the shortcut. (Until we encounter evidence of noninvariance in the difference tests.)

At the beginning of the model results section of a shortcut analysis, we will find overall fit information for the three multiple group models and information on changes in fit that occurred as loading constraints and then threshold constraints were added.

Note: Some researchers and the online Mplus User's Guide suggest evaluating only the Configural/Scalar model comparison (Millsap & Yun-Tein, 2004). They argue that loadings and thresholds for an indicator should only be freed or constrained in tandem. With this approach, the code "MODEL IS CONFIGURAL SCALAR;" would be used, and searches for sources of noninvariance would involve freeing each indicator's loadings and thresholds simultaneously. Freeing and constraining loadings and thresholds simultaneously could also be implemented with the traditional approach.

Because there are advantages of testing factor loadings and thresholds separately (Bowen & Masa, 2015; Dimitrov, 2010), we include the metric model in our examples.

A shortcut analysis generates substantially more output than usual because after the overall fit and comparison output, the usual detailed fit and parameter output is presented for all three models and both groups. With so much output, it is critical to pay attention to which model and which group the output is referring to in these examinations.

SHORTCUT STEP 1: Test your hypothesized CFA model separately with each group.

The first step of the shortcut approach is the same as Step 1 of the more traditional approach. Carefully read text on Step 1 that starts at bottom of page 2 of this handout.

SHORTCUT STEP 2: Specify a multiple group model based on results from the individual group CFAs.

The second step of the shortcut testing sequence encompasses Steps 2, 3, 4 and 6 from the traditional approach. The shortcut generates and compares fit (as measured by χ^2) of the configural, metric, and scalar models all in one step. The “MODEL IS: CONFIGURAL METRIC SCALAR;” statement is put in the ANALYSIS section of the input file. The shortcut input for the example we have been using would be:

```
VARIABLES
  NAMES ARE: . . . .
  USEVARIABLES ARE  wb1 wb2 wb3 wb4 wb5;
  CATEGORICAL ARE  wb1 wb2 wb3 wb4 wb5;
  MISSING ARE ALL (99);
  GROUPING IS RACETH (1 = AA 3 = WH );
ANALYSIS:
  ESTIMATOR IS WLSMV;
  MODEL IS CONFIGURAL METRIC SCALAR;
MODEL:
  WELLBEING by wb1 wb2 wb3 wb4 wb5;
MODEL WH:
```

The unboxed top section of the Invariance Testing output below indicates the configural model has good fit (non-significant χ^2). χ^2 gets larger as more constraints are added. Both the metric and scalar models have statistically significant χ^2 s. We expect fit to get worse as we add constraints so those results are not surprising.

The question of most interest is: does fit become **significantly** worse with added constraints. To answer that question, we look at the output in the box. The χ^2 for the Metric against Configural comparison is statistically significant. Like the output obtained from Step 4 in the traditional approach, this result indicates there is noninvariance across the two groups among the loadings. The χ^2 for the Scalar against Metric comparison is **not** statistically significant. This too is consistent with what we found in Step 6 of the traditional approach: the thresholds are invariant. In fact, the χ^2 values in the box are the same as those found in the traditional model testing.

Mplus Shortcut Output for Shortcut Step 2

MODEL FIT INFORMATION

Invariance Testing

Model	Number of Parameters	Chi-Square	Degrees of Freedom	P-Value
Configural	24	37.889	10	0.0630
Metric	21	56.725	14	0.0032
Scalar	18	59.581	18	0.0022

Models Compared	Chi-Square	Degrees of Freedom	P-Value
Metric against Configural	20.422	4	0.0004
Scalar against Configural	23.220	8	0.0031
Scalar against Metric	4.292	4	0.3679

SHORTCUT STEP 3: Explore Partial Invariance Among Loadings and/or Thresholds if indicated by the results of Shortcut Step 2.

If indicated by the initial shortcut output, Steps 5 and 7 of the traditional approach can be used in conjunction with the shortcut to identify the sources of noninvariance among loadings and/or thresholds and determine if partial invariance is tenable. Even when it is necessary to explore partial invariance using strategies from traditional Steps 5 and 7, the shortcut makes the process easier. Below we describe the four scenarios possible in the results from Shortcut Step 2 and recommend how to specify follow-up shortcut analyses given each scenario.

Note: It might seem logical to test one **freed** loading or threshold at a time when using the Shortcut. However, from experience and consistent with (Brown, 2015), we recommend testing one **constrained** loading or threshold at a time. Modification indices, which might help us decide which parameters to free first, are not available with the shortcut. It is also easy (for some of us) to make mistakes interpreting findings when freeing one parameter at a time. Brown provides other reasons. The procedures below follow the one-constraint-at-a-time logic. This approach is also consistent with the manual difftest (traditional) approach, in which the less restrictive model has to be tested first.

Four Shortcut Scenarios

Four scenarios are possible when we examine results for the Metric against Configural and Scalar against Metric comparisons. In our example above, there is noninvariance among the loadings but not among the thresholds. This situation is discussed in Scenario 2. Other possible results from the initial shortcut analysis and how to handle them are described in Scenarios 1, 3, and 4.

Scenario 1: The p values for both comparisons are greater than .05.

Non-significant χ^2 s in the shortcut output for both the Metric against Configural, Scalar against Metric (or Scalar versus Configural) comparisons are evidence of scalar invariance (which includes metric invariance). The results indicate that constraining all loadings and then all thresholds across groups did not significantly worsen fit. We would happily conclude the construct is measured equivalently and means the same thing across the two groups. After establishing scalar invariance, we would evaluate the fit indices and parameter estimates for the final scalar model. In addition to scalar invariance, we need evidence of good fit and acceptable parameters for the scalar model before we start using the latent variable in analyses with other variables.

Scenario 2: The p value for the Scalar against Metric comparison model is greater than .05, but for the Metric against Configural comparison, p is less than .05.

Scenario 2 is what we have in the output shown in Step 2. We conclude all thresholds are invariant, but one or more of the factor loadings is noninvariant. Our next step would be to constrain just one factor loading at a time and re-examine the shortcut output related to the Metric against Configural model. The hope is we can identify a partially invariant metric model—one in which 20% or fewer of the loadings are non-significant (Dimitrov, 2010).

We would essentially follow the guidelines in Step 5 of the traditional approach, but with important shortcut assistance. Specifically, the shortcut automatically specifies configural

and metric models and conducts the difftest, so we have less code to enter than we did in the traditional approach. The shortcut configural model is the comparison model we previously had to create by freeing each freeable loading and threshold. The default shortcut metric model has loadings constrained to be equal and free thresholds, also a model we had to create with several lines of code in the traditional approach.

As with the traditional approach, default shortcut models can be modified by statements in the MODEL WH: section of the input. Modifications related to loadings will alter the metric model, and the modified model will be used in the automatic difftest analyses. The configural model will stay the same. As with the traditional approach, we take the approach of testing the effects on the metric against configural comparison of adding one loading constraint at a time. As with the traditional approach, we free all but the loading we are testing (and the referent loading). By default the metric model is specified with free thresholds; we do not have to list freeable thresholds as we did in the traditional approach.

The relevant code for testing just one loading constraint at a time is presented below. The input commands below will generate Model Comparison data for the configural model against a metric model with only the loading for wb2 constrained to be equal across groups.

ANALYSIS: MODEL IS CONFIGURAL METRIC;	No need for all the SCALAR output right now.
MODEL: WELLBEING BY wb1 wb2 wb3 wb4 wb5;	The model for both groups has 1 factor and 4 indicators. Wb1 is the referent.
MODEL WH: WELLBEING BY wb3 wb4 wb5;	We override the METRIC model defaults by repeating part of the WELLBEING statement in the WH group. Wb3, wb4, and wb5 are freed. Wb1 remains fixed; Wb2 is not listed, so is constrained. Its effects on fit are being tested.

As in the traditional approach, we record the χ^2 and p value for the Metric against Configural comparison generated in the analysis. For example:

Constrain only wb2: chi sq = 17.2, p = .0000

The result matches what was found in the traditional approach. Now we alter the MODEL WH: statement so only the next loading, wb3, is constrained to be equal across groups; steps (wb2, wb4 and wb5 are free, wb1 remains fixed to 1).

MODEL WH:

WELLBEING BY wb2 wb4 wb5;

Constrain only wb3: chi sq diff = 2.8, p = .0942

And we repeat the steps for wb4 and wb5 and obtain the following results.

Constrain only wb4: chi sq diff = 6.7, p = .0165
 Constrain only wb5: chi sq diff = 4.7, p = .0309

This is the same sequence of tests used in the traditional approach, but simpler to execute. The results are the same as with the traditional approach: wb2, wb4, and wb5 are noninvariant. Wb3 is the best choice for referent in the test of wb1's invariance, so again we respecify the factor model so wb3 is the referent in both groups.

MODEL:

WELLBEING BY wb3 wb1 wb2 wb4 wb5;

We make wb3 the referent.

MODEL WH:

WELLBEING BY **wb2 wb4 wb5;**

We omit wb1 from the list of freed parameters so we can test its invariance.

The result is again consistent with the traditional testing results: wb1 is invariant:

Constrain only wb1: chi sq diff = 2.8, p = .0942

Based on the shortcut results, we would again conclude the well-being measure has metric noninvariance across the two groups. **If**, however, we had met partial invariance criteria for the loadings, our invariance testing would be complete for this scenario because threshold invariance was already established in the initial shortcut results. We would conclude the measure has partial scalar invariance, and that we can compare the means and variances of scores on the latent variable across groups and can include both groups in analyses with the latent variable. The final essential step would be to evaluate the fit indices and parameter estimates for the final scalar model (with the noninvariant loading modeled) in the shortcut output. In addition to scalar invariance, we need evidence of good fit and acceptable parameters for the scalar model.

Scenario 3: The p value for the Metric against Configural model is greater than .05, but the p value for the Scalar against Metric comparison is less than .05.

With this scenario, we conclude the factor loadings are invariant, but one or more thresholds is noninvariant. Our next step is to constrain one threshold at a time and re-examine the shortcut χ^2 diff test output related to the Scalar against Metric comparison.

ANALYSIS:

MODEL IS METRIC SCALAR;

We are focusing on the scalar-metric comparison so we don't need all the CONFIGURAL output.

MODEL:

WELLBEING BY wb1 wb2 wb3 wb4 wb5;

The model for both groups has 1 factor and 5 indicators. Wb1 is the referent indicator.

MODEL WH:

[wb3\$2];

[wb4\$2];

[wb5\$2];

We override the SCALAR model defaults (all thresholds constrained) by naming 3 of 4 freeable thresholds, which will result in them being freed.

Wb2\$2 will be tested for invariance because its omission here means it remains constrained.

We can't test either threshold for the referent indicator—they both have to be constrained—so we'll start with wb2's threshold. For identification, all \$1 thresholds are constrained to equality. Mplus will know the scalar model is being modified because our MODEL WH: statement involves thresholds. The metric model will not be affected.

The scalar-metric comparison will now be between the metric model and a scalar model with only one threshold constrained. After testing and recording the effects of constraining wb2\$2, we could test wb2\$1 by adding it to the list of freed thresholds instead of wb2\$2. Mplus would then automatically constrain wb2\$2 to equality and test the effects of constraining wb2\$1 on fit relative to the metric model. We would proceed to test all the thresholds, including those for wb1. As before, we can test wb1\$1 and wb1\$2 by making another variable the referent. As with loadings, if more than two thresholds are noninvariant in individual parameter tests, we could free the two **most** noninvariant ones, and see if any other less noninvariant threshold performs better when tested after more noninvariant parameters are freed. These steps are the same as those outlined in Step 7 of the traditional approach and will yield the same results.

Scenario 4: The p values for both comparisons are less than .05.

With this least desirable scenario from the initial CONFIGURAL METRIC SCALAR TEST, there is noninvariance in both the metric and scalar models. To determine if partial invariance pertains, we would follow the steps for Scenario 2 (exploring partial invariance in the metric model) and then, if the partial metric invariance is found, steps for Scenario 3 (exploring partial invariance among thresholds). If, as in our original model, metric invariance is not demonstrated, we stop testing and conclude the measure is noninvariant. When proceeding to scalar model tests, the final partially invariant metric model is specified as the comparison model. For example, if one loading needed to be free in order to have a non-significant metric-to-configural model comparison, we would keep that loading free as we test for partial invariance in the scalar model. If fewer than 20% of parameters in each model are noninvariant, and the final scalar model has adequate fit and parameter estimates, we can claim partial invariance for the measure and include both groups in subsequent SEMs with the latent variable. In those analyses the shortcut code would be removed, but the grouping variable specification would remain. Partial invariance would be modeled using the conventional code (which specifies the default scalar model) along with code for the small number of freed parameters identified in the invariance testing. For example:

```
VARIABLES . . . .
    GROUPING IS RACETH (1 = AA 3 = WH );
ANALYSIS: . . .
    !MODEL IS CONFIGURAL METRIC SCALAR;
MODEL:
    WELLBEING by wb3 wb2 wb1 wb4 wb5;
    Structural model specification
MODEL WH:
    WELLBEING by wb2;
    [wb3$2];
```

References

- Bauer, D. J. (2017). A More General Model for Testing Measurement Invariance and Differential Item Functionin. *Psychological Methods*, 22(3), 507-526. <https://doi.org/10.1037/MET0000077>
- Bollen, K. A. (1989). *Structural equations with latent variables*. John Wiley & Sons.
- Bowen, N. K. (2021a). Brief Reference: Chi Square Difference Testing in Mplus with WLSMV Estimation
- Bowen, N. K. (2021b). *Reference Graphic: Mplus Specification of Levels of Measurement Invariance with Ordinal Data*
- Bowen, N. K. (2021c). *Requesting and Using Modification Indices in Mplus*.
- Bowen, N. K., & Masa, R. D. (2015). Conducting measurement invariance tests with ordinal data in Mplus: A guide for social work researchers. *Journal of the Society for Social Work and Research*, 6(2), 1-21. <https://doi.org/10.1086/681607>
- Brown, T. A. (2015). *Confirmatory factor analysis for applied research, 2nd Edition*. Guilford Press.
- Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management*, 25(1), 1-27.
- Dimitrov, D. M. (2010). Testing for factorial invariance in the context of construct validation. *Measurement and Evaluation in Counseling and Development*, 43(2), 121-149. <https://doi.org/10.1177/0748175610373459>
- Lai, K. (2020). Fit Difference Between Nonnested Models Given Categorical Data: Measures and Estimation. *Structural Equation Modeling: A Multidisciplinary Journal*, 1-22. <https://doi.org/10.1080/10705511.2020.1763802>
- Liu, Y., & Sriutaisuk, S. (2021). A Comparison of FIML- versus Multiple-imputation-based methods to test measurement invariance with incomplete ordinal variables. *Structural Equation Modeling: A Multidisciplinary Journal*, 1-19. <https://doi.org/10.1080/10705511.2021.1876520>
- Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Journal of Multivariate Behavioral Research*, 39, 479–515-479–515.
- Muthén, L. K., & Muthén, B. O. (2019). *Mplus (version 8.4)*. In Muthén & Muthén.
- Muthén, L. K., Muthén, B. O. (1998-2017). *Mplus User's Guide, 8th Ed*. Muthén & Muthén.
- Sass, D. A. (2011). Testing measurement invariance and comparing latent factor means within a confirmatory factor analysis framework. *Journal of Psychoeducational Assessment*, 29(4), 347-363. <https://doi.org/10.1177/0734282911406661>

- Sass, D. A., Schmitt, T. A., & Marsh, H. W. (2014). Evaluating Model Fit With Ordered Categorical Data Within a Measurement Invariance Framework: A Comparison of Estimators. *Structural Equation Modeling: A Multidisciplinary Journal*, 21(2), 167-180.
<https://doi.org/10.1080/10705511.2014.882658>
- Wu, H., & Estabrook, R. (2016, Dec). Identification of Confirmatory Factor Analysis Models of Different Levels of Invariance for Ordered Categorical Outcomes. *Psychometrika*, 81(4), 1014-1045.
<https://doi.org/10.1007/s11336-016-9506-0>