Chemistry
Education Research
and Practice

Addressing diversity and social inclusion through group comparisons: A primer on measurement invariance testing

Journal:	*Chemistry Education Research and Practice*
Manuscript ID	RP-ART-01-2020-000025.R1
Article Type:	Paper
Date Submitted by the Author:	24-Mar-2020
Complete List of Authors:	Rocabado, Guizella; University of South Florida, Chemistry Komperda, Regis; San Diego State University, Chemistry & Biochemistry Lewis, Jennifer; University of South Florida, Department of Chemistry; University of South Florida, Center for the Improvement of Teaching and Research on Undergraduate STEM Education Barbera, Jack; Portland State University, Department of Chemistry

Addressing diversity and social inclusion through group comparisons: A primer on measurement invariance testing

Guizella A. Rocabado,[1] Regis Komperda,[2‡] Jennifer E. Lewis[1,3] and Jack Barbera[4†]

Abstract

As the field of chemistry education moves toward greater social inclusion, social justice, and increased participation by underrepresented minorities, standards for investigating the differential impacts and outcomes of learning environments have to be considered. While quantitative methods may not be capable of generating the in-depth nuances of qualitative methods, they can provide meaningful insights when applied at the group level. Thus, when we conduct quantitative studies in which we aim to learn about the similarities or differences of groups within the same learning environment, we must raise our standards of measurement and safeguard against threats to the validity of inferences that might favor one group over another. One way to provide evidence that group comparisons are supported in a quantitative study is by conducting measurement invariance testing. In this manuscript, we explain the basic concepts of measurement invariance testing within a confirmatory factor analysis framework with examples and a step-by-step tutorial. Each of these steps is an opportunity to safeguard against interpretation of group differences that may be artifacts of the assessment instrument functioning rather than true differences between groups. Reflecting on and safeguarding against threats to the validity of the inferences we can draw from group comparisons will aid in providing more accurate information that can be used to transform our chemistry classrooms into more socially inclusive environments. To catalyze this effort, we provide code in the electronic supplementary information (ESI) for two different software packages (R and Mplus) so that interested readers can learn to use these methods with the simulated data provided and then apply the methods to their own data. Finally, we present implications and a summary table for researchers, practitioners, journal editors, and reviewers as a reference when conducting, reading, or reviewing quantitative studies in which group comparisons are performed.

[1] Department of Chemistry, University of South Florida.

[2] Department of Chemistry and Biochemistry; Center for Research in Mathematics and Science Education, San Diego State University.

[3] Center for the Improvement of Teaching and Research in Undergraduate STEM Education

[4] Department of Chemistry, Portland State University.

[†] Corresponding author for manuscript (jbarbera@pdx.edu)

[‡] Corresponding author for ESI (rkomperda@sdsu.edu)

Electronic Supplementary Information (ESI) available: [ ] See DOI:

Introduction

Diversity and social inclusion are popular terms in science education at present. In the past few decades, numerous research endeavors have focused on studying diverse populations of students within science, technology, engineering, and mathematics (STEM; e.g., Hong and Page, 2004; Tsui, 2007; Hurtado *et al*., 2010). Due to a directive to increase minority representation in STEM fields in the United States (Seadler, 2012), colleges and universities have launched initiatives to attract underrepresented minority (URM) students. These initiatives can help to initially increase diversity representation; however, simply admitting students is not enough if they feel unvalued or unwelcome in their college communities (Puritty *et al.,* 2017). Thus, diversity initiatives may fail to retain these students without attention to creating inclusive environments where students of all backgrounds feel they have a voice and that they matter (Puritty *et al.,* 2017). Attaining a diverse STEM workforce, then, means promoting social inclusion and social justice in our classrooms and in our research (O'Shea *et al.,* 2016).

Critical Race Theory (CRT) has become a central framework to study issues of inclusion and social justice, particularly for members of marginalized racial groups (Crenshaw, 1995; Solórzano, 1997, 1998; Delgado and Stefanic, 2001; Yosso, 2005; Dixson and Anderson, 2018). Although CRT was born in the legal realm, it has permeated the educational field as well (Crenshaw, 1995; Delgado and Stefanic, 2001). This theory has been linked to five guiding tenets that inform research, curriculum, pedagogy, and policy (Solórzano, 1997; Yosso, 2005). Three of these tenets seem particularly well suited to investigations utilizing quantitative methodology. First, an acknowledgment of the centrality of race and racism in the power relations that underpin society requires that race be explicitly considered rather than ignored in educational research. Second, the de facto existence of 'dominant ideology' informed by race and racism requires us to cast aside naive beliefs that research and researchers are neutral and objective (Yosso, 2005) and work to safeguard against systemic biases and the propagation of social inequities in educational research (García, López, and Vélez, 2017; Gillborn *et al*., 2018). And third, answering CRT's call for a commitment to social justice requires us to privilege research that works to uncover social inequities and moves toward the eradication of racial and other forms of marginalization (Solórzano, 1997). CRT is a framework well equipped to investigate issues of racism and social inequities in educational settings at the individual as well as at the institutional level. For example, Fernández (2002) uses CRT as a framework and takes an individual approach to display a successful educational experience of one immigrant Latino student in a public school in Chicago via qualitative methods. On the other hand, Solórzano and Ornelas (2004) use CRT to investigate the access and availability of Advanced Placement (AP) courses in California high schools and how they affect African American and Latina/o students' admission to college. This quantitative study exhibits an institutional approach that documents cumulative impacts on individuals and groups of students from minority racial and ethnic populations. Likewise, CRT and quantitative methods can be utilized at the institutional level to investigate achievement gaps in educational systems, providing a wider lens for these investigations (García, López, and Vélez 2017; López *et al.,* 2018), rather than merely grade comparisons. Whenever possible, studies of this nature benefit from a comprehensive investigation with appropriate categories for investigating achievement gaps, such as race-gender-class intersections (Crenshaw, 1989; Covarrubias, 2011, 2013; Litzler, Samuelson and Lorah, 2014; García, López, and Vélez, 2017; Ireland *et al.,* 2018; López *et al.,* 2018) as a movement to achieve a more complete view of the investigation and avoid reproduction of

widespread inequities in educational settings (García, López, and Vélez 2017; Gillborn *et al.,* 2018).

In an effort to combat against racism and other societal inequities, these issues have long been studied with qualitative methodologies (Gillborn *et al.,* 2017; García, López and Vélez, 2018). Quantitative methods have been criticized for an inability to speak to the details of lived experiences of diverse populations (García, López, and Vélez, 2018) and thus been deemed inappropriate to study these issues in educational settings due to these everyday experiences having deep roots in social relationships (Apple, 2001). Although qualitative methods are more appropriate to capture nuances of societal processes as experienced by individuals, quantitative methods can explore wider structures in which individual and collective experiences are lived, revealing wider structural issues that affect these diverse groups on a larger scale (Gillborn *et al.,* 2017). With this tension between qualitative and quantitative methodologies attending to issues of social inequities, we encourage the use of either or both types of methods when appropriate, following the tenets of CRT. Therefore, in an effort to promote inclusion and equity in our classrooms, appropriate qualitative and quantitative methods can be used in research, with the premise that our methods must be reflexive and safeguarded against systemic racial, ethnic, gender, and other biases favoring the majority groups (Gillborn *et al.,* 2017).

Much of the critique about using quantitative methods to investigate these issues comes from the problem that numbers are positioned as 'neutral' and audiences may believe 'data speaks for itself.' Critical theorists argue that these claims of neutrality are far from the truth (Gillborn *et al.,* 2017). However, researchers, practitioners, and policy-makers tend to put great emphasis in numbers, as these are the data by which policies are justified and schools and districts are labeled successes or failures (Gillborn *et al.,* 2017). Thus, to rise above these critiques in favor of continuing to use quantitative approaches to investigate social inequities, a process of ongoing self-reflexivity and engagement with historical, social, and political structures of the groups under investigation must be present (García, López and Vélez, 2018). Additionally, because numbers carry such important consequences, we must use them with caution and systematically interrogate the validity of the inferences we make with these numbers, particularly as it relates to consequential validity (AERA, NCME and APA, 2014). According to Messick (1995) the social consequences of score interpretation may be positive or negative, intentional or unintentional. Thus, in the interest of advancing inclusion and social justice, researchers must engage in collecting evidence of positive consequences while minimizing adverse effects. As an example of unintentional, negative effect, one could imagine that a subgroup of students misinterprets items on an assessment instrument based on unfamiliar words in the item, which may lead to confounding results in the data for that subgroup. This source of invalidity can potentially lead to erroneous decisions that may have adverse consequences for this subgroup of students (Shephard, 1993; Messick, 1995). Therefore, raising the bar for quantitative methods in our field will require taking steps to safeguard against consequential validity threats that may be present when making group comparisons.

Quantitative Standards for Group Comparisons in CER

In CER, investigations of efforts to broaden participation of diverse student populations have been a focus of multiple studies (i.e., Richards-Babb and Jackson, 2011; Rath *et al.,* 2012; Fink *et al.,* 2018; Stanich *et al.,* 2018; Nawarathne, 2019; Shortlidge *et al.,* 2019). Many of these

studies have aimed to investigate differential outcomes of URM students by performing group comparisons with various statistical analyses (Rath *et al.,* 2012; Fink *et al.,* 2018; Stanich *et al.,* 2018; Shortlidge *et al.,* 2019). For instance, Fink and colleagues (2018) proposed a strategy to promote improved general chemistry performance for women and minorities through a growth mindset intervention. The results of the study report higher performance overall favoring the White students; however, post-hoc Tukey tests confirmed an intervention effect for minority students, who ultimately earned more than 5 percentage points higher on average in the mindset intervention condition (Fink *et al.,* 2018). Similarly, Stanich and colleagues (2018) implemented a supplementary instruction (SI) course that aimed to narrow achievement gaps by showing that URM students who participated in the SI course had lower failure rates in general chemistry than URM students who did not take the course. Additionally, this study also aimed to narrow affect gaps by increasing perception of relevance, sense of belonging, and emotional satisfaction toward the subject of chemistry (Stanich *et al.,* 2018). While studies such as these are a positive sign that diversity and social inclusion are being taken seriously, there is still work to be done with respect to developing guidelines for quantitative research on these issues.

The next important step in developing research standards is to critically examine the collection, analysis, and representation of quantitative data and results for threats to the validity of inferences when group comparisons are to be made. CER has a long history of assessment design to probe student understanding of concepts taught in the classroom (i.e., Tobin and Capie, 1981; Roadrangka, Yeany and Padilla, 1983; Loertscher, 2010; Villafañe, *et al.,* 2011; Kendhammer, Holme and Murphy, 2013; Wren and Barbera, 2013; Brandriet and Bretz, 2014; Bretz, 2014; Kendhammer and Murphy, 2014; Xu, Kim and Lewis, 2016). These, and other, assessment instruments have been used by researchers and practitioners to evaluate the success of classroom interventions and curricular changes. Furthermore, in the last few decades, CER as a field has moved toward an increased interest in affect and motivation in educational settings (Xu, Villafañe and Lewis, 2013; Ferrell and Barbera, 2015; Salta and Koulougliotis, 2015; Ferrell, Phillips and Barbera, 2016; Liu *et al.,* 2017; Gibbons and Raker, 2018; Gibbons, *et al.,* 2018; Hensen and Barbera, 2019; Rocabado *et al.,* 2019). Thus, assessment instruments may be used in CER to determine research agendas, report findings, evaluate interventions or curricular design and much more.

Given the current interest in measuring affect in the classroom, there is an added concern that many cognitive and emotional factors might have different effects among diverse populations, particularly disfavoring URM groups (Ceci, Williams and Barnett, 2009; Villafañe, García and Lewis, 2014; Rocabado, *et al.,* 2019). However, some of the differences noted in these data could be an artifact of the assessment instrument (Jiang, García and Lewis, 2010); thereby resulting in a potential threat to the validity of the inferences drawn from the instrument-derived data (Arjoon, Xu and Lewis, 2013; AERA, APA and NCME, 2014). Therefore, in the interest of promoting social inclusion in the classroom, it is important to know that when an instrument functions well for the whole class, the functionality extends to any subgroups of interest. Nevertheless, simply comparing observed scores for subgroups is not appropriate. As shown by several studies (Khaveci, 2015; Komperda, Hosbein and Barbera, 2018; Montes, Ferreira and Rodriguez, 2018), differences might arise as artifacts of the instrument functioning and not as differences in understanding, ability, or affect.

Goals of This Measurement Invariance Testing Primer

To encourage and support the gathering of evidence to substantiate group comparisons within CER, this manuscript presents the quantitative method of measurement invariance testing for those familiar with factor analysis. A comprehensive review of measurement invariance testing can be found in Vandenberg and Lance (2000). Measurement invariance testing can be used to investigate the degree to which measured student data is represented by the same theoretical model. Prior to introducing the details and meanings of the various levels of measurement invariance testing, we discuss latent variables and data visualization techniques. This introduction provides initial insight into the relations among assessment items as well as providing a basis for understanding the mathematical foundations being tested. We then provide a step-by-step tutorial of measurement invariance testing, discussing what is being tested, how to evaluate if invariance has been achieved, and what (if any) comparisons between groups are supported at each step. Finally, we present a summary of the implications of measurement invariance testing as well as recommendations for researchers, practitioners, reviewers, and journal editors.

Group Comparisons on Latent Constructs

Commonly, the variables of interest in CER are ones that cannot be measured directly, i.e., they are latent traits. Variables such as student self-efficacy, attitude, metacognition, mindset, and understanding of chemistry are all examples of latent traits. Many of these latent traits are multidimensional, that is, they are subdivided into smaller latent units (subconstructs or factors) that make up the latent trait (Brown 2006 pp.2). To provide an example for our discussion of quantitative data comparison by group, we devised a *fictitious* assessment instrument to measure the latent trait of 'perceived relevance' toward chemistry. Such an instrument might be useful in understanding college students' perceptions of the field of chemistry. For this *fictitious* assessment instrument, it could be expected that students' perceived relevance of chemistry might differ by college major and that a researcher might want to compare data from this instrument by group. While many times the groupings of students we quantitatively investigate are by gender or URM status, these are not the only groupings for which comparisons need to be supported by evidence. For example, with our *fictitious* instrument the comparison groups could be defined as STEM and non-STEM majors. Other groupings could be first-generation college students or community college transfer students for comparison to students not in these groupings. Whatever the chosen comparison groups are, it is imperative that researchers have a directive to investigate those groups and use an appropriate construct for the comparison.

It is important to note that utilizing assessment instruments that have been developed with a strong theoretical background and which have been investigated for forms of validity and reliability evidence delineated by the *Standards for Educational and Psychological Testing* (Arjoon, Xu and Lewis, 2013; AERA, APA and NCME, 2014) is imperative to drawing meaningful insights from studies. Following with the example, and assuming that the instrument was created under these conditions, our *fictitious* assessment instrument is called the Perceived Relevance of Chemistry Questionnaire (PRCQ) and contains three *fictitious* subconstructs: Importance of Chemistry (IC), Connectedness of Chemistry (CC), and Applications of Chemistry (AC). The *fictitious* PRCQ is a 12-item instrument with four items per subconstruct. When student responses to these 12 items are examined, the expected pattern of bivariate

correlations among responses would be that items aligned with the same subconstruct should have stronger correlations with each other, meaning they are highly associated with each other through an underlying subconstruct, and have weaker correlations with other items aligned with different subconstructs. For comparison purposes then, these item-level patterns need to be consistent within each group.

Group Comparisons Through Data Visualization

In addition to using descriptive statistics to investigate data patterns, item-level data can be visually inspected using a variety of methods (e.g., box-plots, violin plots, graphs, charts). To demonstrate ways in which to visualize data, we have created simulated PRCQ datasets that highlight several different data patterns across groups (see ESI pp. 1-7 for additional details). Item correlation values for one of these datasets are plotted in a correlation heatmap shown in Figure 1. In this correlation plot the item labels (i.e., I1, I2, etc.) are listed on the diagonal, and the color of each square represents the value for the correlation (i.e., the strength of association) between two items. Pairs of items with stronger correlations are represented with darker squares and pairs of items with weaker correlations are represented with lighter squares. The simulated data used in this example are strongly correlated in four-item sets (I1 to I4, I5 to I8, and I9 to I12); items outside these sets (e.g., I1 and I8) are weakly correlated. As the PRCQ has three subconstructs, another way to represent the relations between the twelve items is with a factor diagram. The intended factor diagram for the 12-item PRCQ instrument has been added above the correlation plot. In a factor diagram each individual item (called an indicator item and represented by a square) is associated with a subconstruct or factor (represented by a circle). Together, these visual representations of the PRCQ data provide initial visual evidence for the presence of the intended factors (i.e., item set groupings).

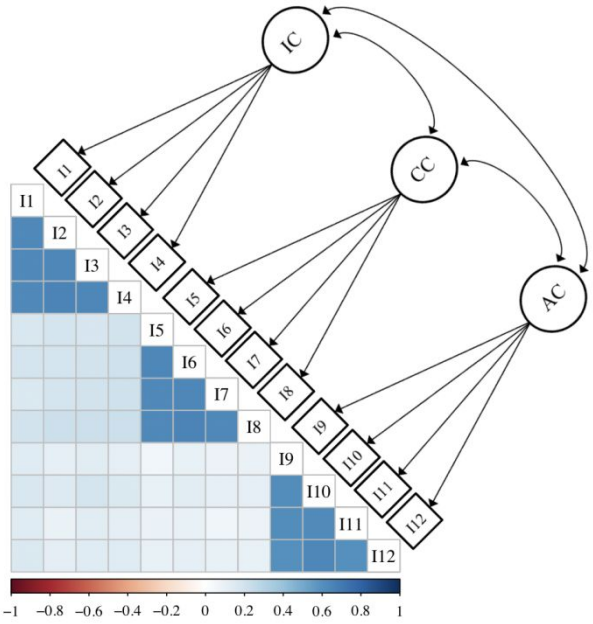


Figure 1. A visualization of the lower correlation matrix for the 12-item PRCQ instrument with a factor model overlaid to illustrate how correlations between sets of items implies the presence of an underlying factor structure. We note that, although the covariance matrix is more directly applicable, the correlation matrix is a standardized covariance matrix, and therefore easier to visualize and discuss.

When making measurements that will ultimately be used to compare the outcomes of various groups on an underlying construct (i.e., Importance of Chemistry (IC), Connectedness of Chemistry (CC), and Applications of Chemistry (AC)), it is necessary to provide evidence that the PRCQ instrument is functioning in a similar way for each group being compared. This practice is a way in which the field of CER can meet best practices when making comparisons and provide evidence to support that any differences between the groups' data are due to true differences in the construct, not a result of systematic bias in the measurement of the construct (Gregorich 2006; Sass 2011). Using our example, as researchers we could be interested in measuring potential differences in the perceived relevance of chemistry (as measured by the PRCQ) between groups. As lower-level chemistry courses serve a range of majors, we could investigate potential differences in perceived relevance between STEM and non-STEM majors, or among multiple groups such as White, African-American, Asian, and Hispanic students. For simplicity in our example, we have simulated response data for a two-group comparison, which will help us visualize the discussion that will proceed. In addition, the data we have simulated is continuous. However, we do understand that much of the data generated in CER is categorical in nature and as such will necessitate a different set of considerations. Thus, we provide explanation and analyses for both continuous and categorical data, in the electronic supplementary information (ESI), along with code (in R and Mplus) for generating the data visualizations as well as the additional analysis steps described later in this manuscript.

If the aggregated PRCQ data in Figure 1 were divided by STEM and non-STEM majors, one step towards examining consistent functioning across groups would be to see if the two groups have similar correlation plots. As shown in Figure 2, when visually comparing the correlation plots by group, it can be seen that they are essentially identical. Ways of testing this similarity statistically will be discussed later.

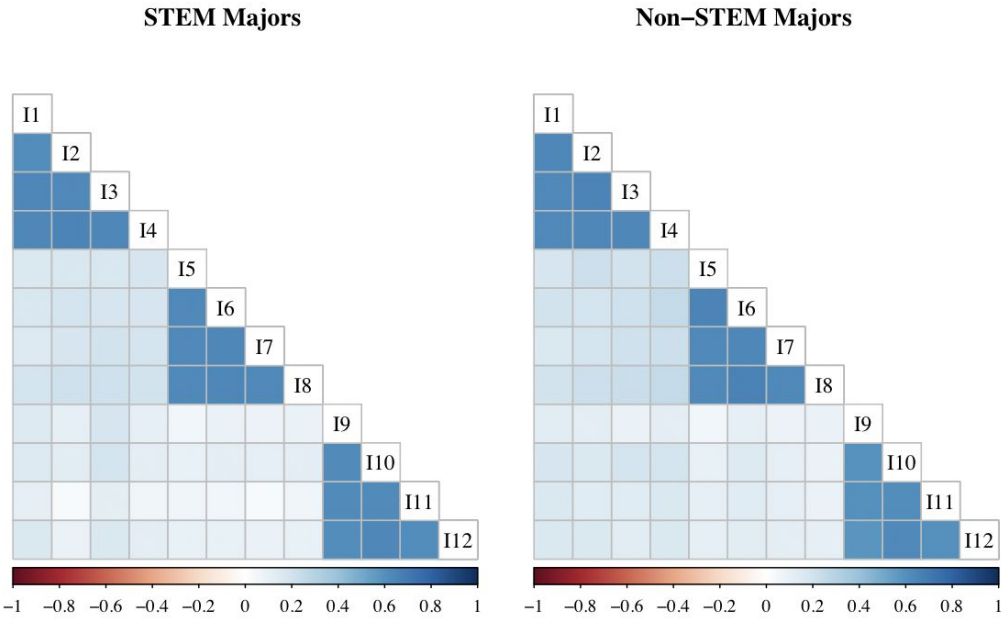


Figure 2. Correlation plots for 12 items with similar strength of association for each item and its intended factor for two subgroups (STEM majors and non-STEM majors) within the data set.

While the situation represented in Figure 2 is the best possible outcome (i.e., the data are simulated to align with a known factor structure for both groups), it is not always the case that data from students in different groups will show the same strength of association between each item and each intended factor. An example of such a situation is visualized in Figure 3 where we simulated a difference in strength of association for one item in one group. In this aggregated PRCQ data set (Figure 3a) we can see inconsistencies around I10, where some correlation boxes are lighter. Although, the overall correlation pattern is consistent (i.e., an instrument that measures three distinct factors as hypothesized for the PRCQ), when we disaggregate the data and view the correlation matrix for each group separately, we observe that I10 has a much lower association with the AC factor for non-STEM majors (Figure 3c) compared to STEM Majors (Figure 3b). This group difference would not be obvious when looking at the correlations in the aggregated dataset (Figure 3a). The situation represented here, dissimilar associations between items and factors across groups, implies that the item is not functioning in similar ways for each group, which could be due to differences in item interpretation for I10. Regardless of the underlying reason, which may never be known for sure, this situation indicates a possible threat to the validity of the potential inferences from the data and needs to be examined more closely to determine whether the data can still be used to compare the groups.

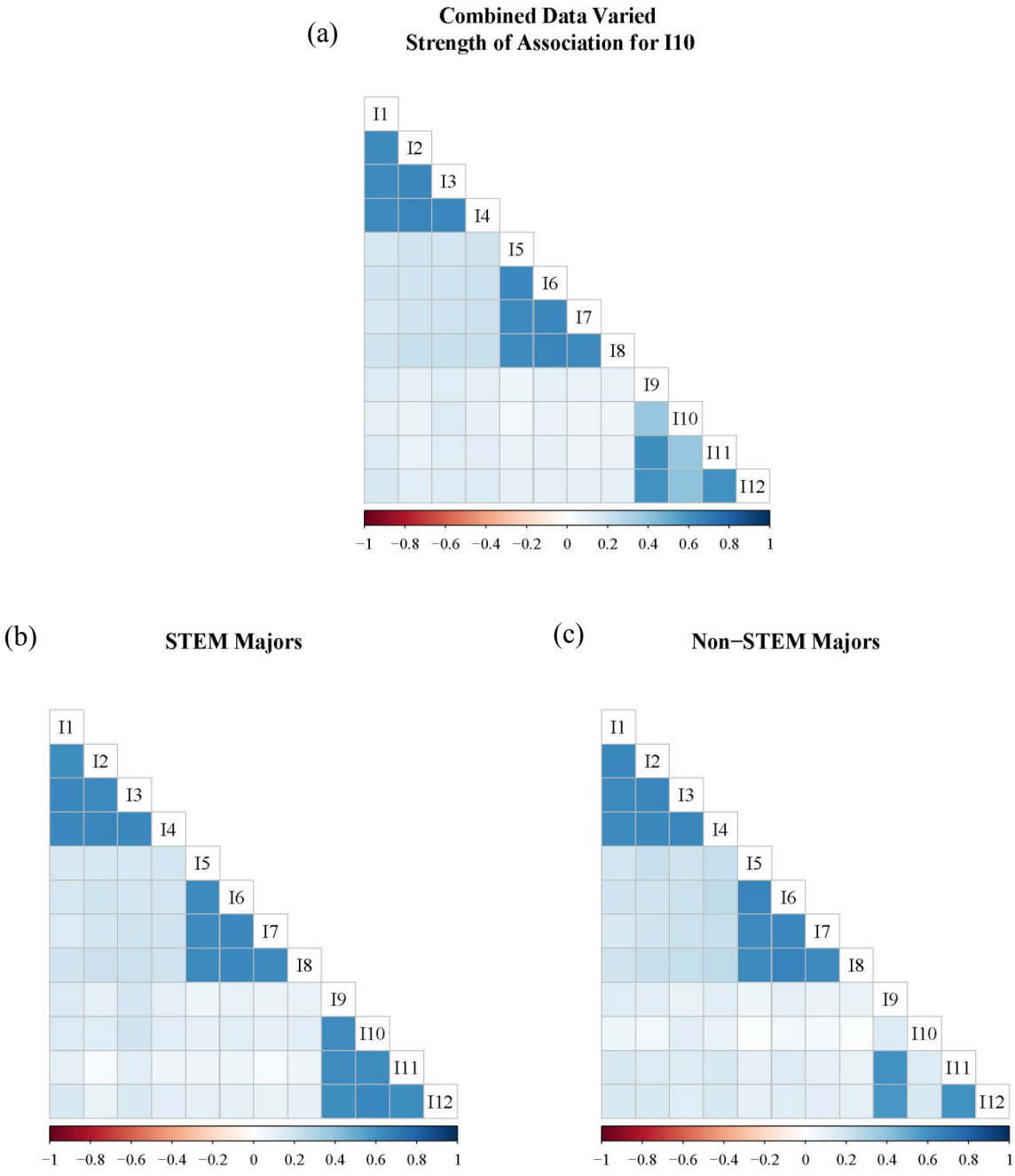


Figure 3. (a) Correlation plot for 12 items with combined dataset; (b) Correlation plot with STEM major data; (c) Correlation plot with non-STEM major data with I10 correlation lowered.

Another type of measurement difference that could occur between the groups is that an item may not have similar response averages in each group. In the next set of simulated data, the strength of association between all items and their intended factor is equivalent, as in Figure 3, but the average response for I3 has been modified for the STEM majors group to illustrate this issue. Unlike when the strength of association differed in the previous example, this result is

more obviously seen when visualizing the correlations in the aggregated dataset (Figure 4a) than in the disaggregated sets (Figures 4b and 4c).

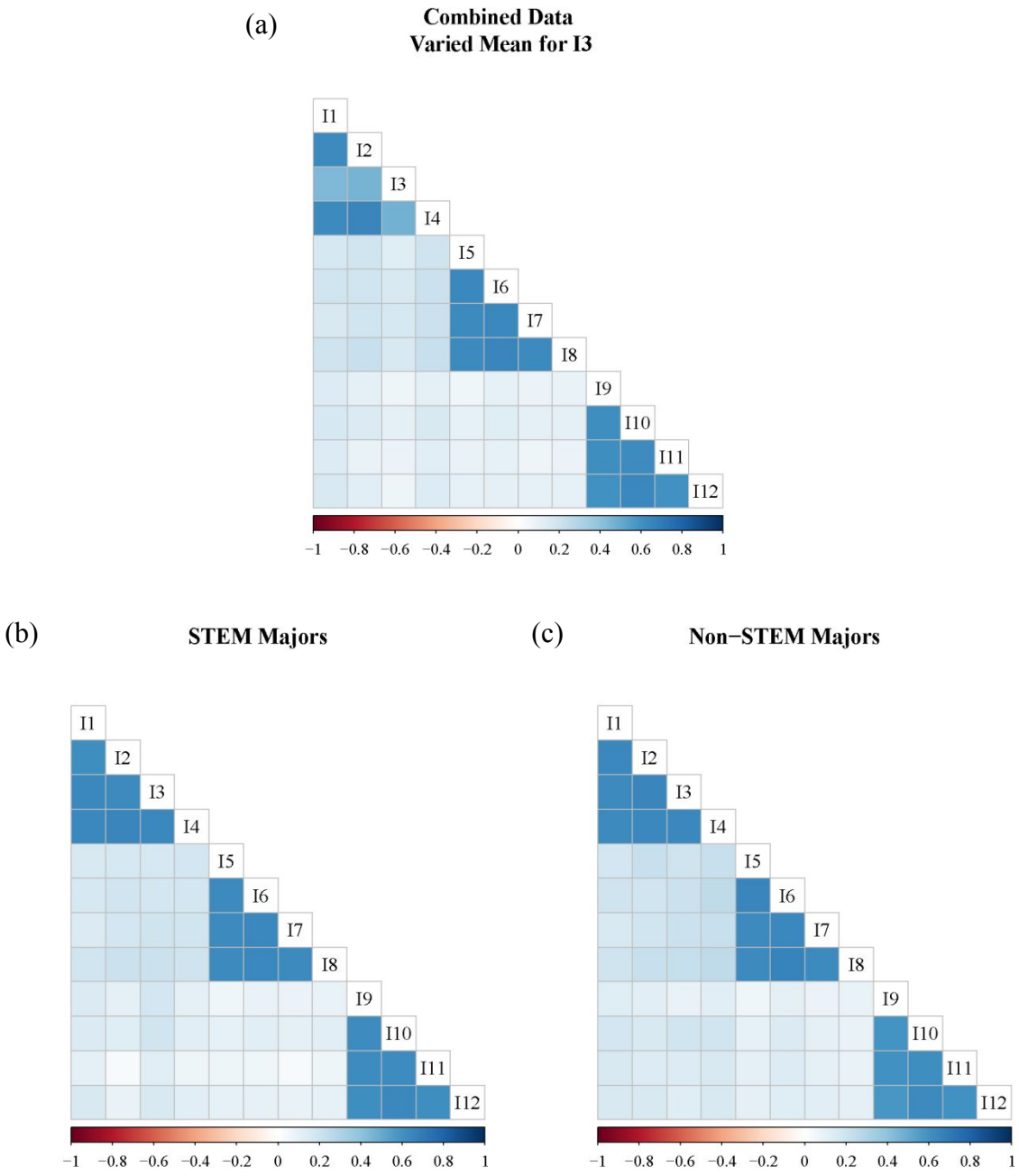


Figure 4. (a) Correlation plot for 12 items with combined dataset; (b) Correlation plot of STEM majors with mean of I3 raised; (c) Correlation plot of non-STEM majors.

To further visualize the distribution of values for each item within each group, Figure 5 plots the means for each item in the two groups using a boxplot. It can be clearly seen that the distribution for I3 in the STEM majors group is much different and is shifted to the higher end of

the scale. This outcome could occur because there are true differences between the groups or it could be due to improper item functioning for one group. However, a quantitative analysis does not differentiate between these two reasons, thus it is appropriate to further investigate the item functioning when this occurs.

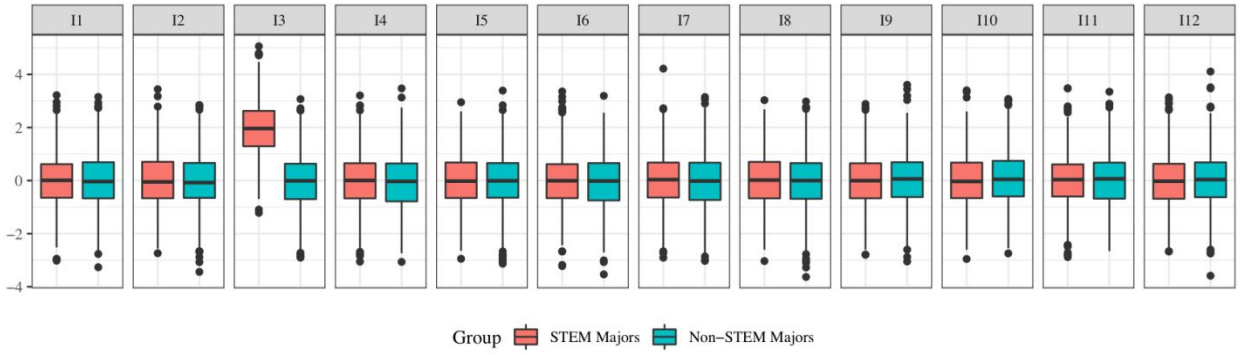


Figure 5. Boxplot of item means for each group.

The item-level differences noted in Figures 3-5 may be due to a variety of issues, which would be worth exploring further in order to understand why they occur. However, in considering whether the data can still be used to make comparisons between groups, the degree to which these differences impact the proposed factor structure need to be evaluated using measurement invariance testing. This quantitative method would indicate if the differences pose a potential issue with how the instrument functions for the different groups, potentially limiting the ability to draw valid conclusions about how the underlying factors of interest differ across groups.

Data Considerations Prior to Performing Measurement Invariance Testing

While we have emphasized the importance of visualizing data and have shown various ways it can be useful, we acknowledge that data visualization is insufficient to address the degree to which item-level differences may impact group comparisons, which necessitates more robust investigations using statistical tests. Additionally, and more often than not, many data issues are not easily visualized, but can become evident in statistical analyses. We encourage all researchers to visualize their data and compute descriptive statistics, thereby providing initial insights to the data as well as evidence about the characteristics of the data. Understanding data characteristics will aid the researcher in making other decisions about further analyses, such as which tests are appropriate to run or which estimator is appropriate to use when modeling data.

Different types of data such as categorical or continuous, can be analyzed with measurement invariance testing utilizing appropriate estimators for each type of data. For example, ordinal data (e.g., categorical data from items with a 7-point Likert-type scale) with variance ranging the entire scale is often treated as continuous data and can be estimated with a maximum likelihood estimator (Muthén and Muthén, 2010; Hirschfeld and von Brachel, 2014). On the other hand, categorical data (e.g., data from a ‘yes or no’ type item or items using fewer than 5 response scale categories) are more appropriately analyzed using a weighted least squares estimator (Muthén and Muthén, 2010; Hirschfeld and von Brachel, 2014; Bowen and Masa, 2015). Ensuring the proper estimator for the data-type is of utmost importance. Violations of

normality, independence, and homogeneity are also important to note, and should be handled appropriately. Discussion of estimators and assumptions is beyond the scope of this article; however, we provide a few resource references for interested readers here (Stevens, 2007; Garson, 2012) and in the ESI.

An additional consideration before conducting measurement invariance testing is statistical power (Hancock and French, 2013). To conduct meaningful statistical analyses, one must ensure an appropriate sample size in order to have enough power to draw meaningful inferences. In measurement invariance testing the interest is in finding no evidence of significant difference between groups, thus, an inappropriate sample size (i.e., too small) can increase the chances of type II error through failing to reject the null hypothesis (of equivalence) when it should have been rejected (Lieber, 1990; Counsell, Cribbie, and Flora, 2019). Recently, work has been done indicating that sample size requirements can be estimated given the number and value of parameters being estimated (Wolf *et al.,* 2013; Mueller and Hancock, 2019).

Confirmatory Factor Analysis Framework

In the previous section we explored visual methods for detecting potential validity threats in our PRCQ data. Though visualizing is an important initial step, more formal statistical methods can and should be employed to evaluate the degree to which differences pose threats to the validity of comparisons. Methods such as Differential Item Functioning have been used to investigate item-level threats in CER (Kendhammer, Holme and Murphy, 2013; Kendhammer and Murphy, 2014), however, the purpose of this paper is to explore threats at the construct, or latent variable level. At this level, various frameworks can be used, including Item Response Theory (IRT; Candell and Drasgow, 1988; Mellenbergh, 1989) and factor analysis (Brown, 2006). As factor analysis methods have become commonplace within CER, and IRT is less frequently utilized in our field, this discussion will focus only on evaluating measurement invariance in a factor analysis framework.

Within a Confirmatory Factor Analysis (CFA) framework (Brown, 2006), measurement invariance testing is a technique that can be used to support that the internal structure of an assessment instrument holds for different groups people at one time point (Salta and Koulougliotis, 2015; Bunce *et al.,* 2017; Hensen and Barbera, 2019; Rocabado *et al.,* 2019) or over time in longitudinal studies (Keefer, Holden and Parker, 2013; Hosbein and Barbera, 2019; Rocabado *et al.,* 2019). In the previous section, the idea of internal structure was described in terms of the grouping of items with each other to form an underlying factor of interest (as introduced in Figure 1). In this section, these associations will be defined more formally using the language of factor analysis.

The CFA framework operates under a network of equations, among which, regression equations link items to latent variables (Brown, 2006). Regression or linear equations (see Equation 1) have several components: a dependent (predicted) variable (y), an independent (predictor) variable (x), the slope of the line (m), the intercept (b), and the measurement error (e).

$$y = mx + b + e \qquad \text{[Eq.1]}$$

Translating the regression equation to the language of factor analysis, the predicted variables are the observed variables (i.e., items), the predictor variables are the factors or latent variables, and the slope is the factor loading. In Figure 6a we write out the regression equation for an item from the PRCQ and in Figure 6b display the model that underlays the PRCQ using common statistical notations in the CFA framework, which we will use for the remainder of the discussion in this manuscript. In this 12-item (i.e., I1-I12), 3-factor (i.e., IC, CC, AC), model lower-case lambdas (λ) represent the factor loading of each item to its respective factor, lower-case taus (τ) represent the intercept of an item, and lower-case epsilons (ε) represent the measurement error of an item. In addition to these parameters, Figure 6b shows the covariance between factors (e.g., double headed arrow between IC and CC) and each individual factor variance (e.g., small curved arrow from IC to IC). While these parameters are part of the overall CFA model for the PRCQ, they do not need to be modified when evaluating for measurement invariance.

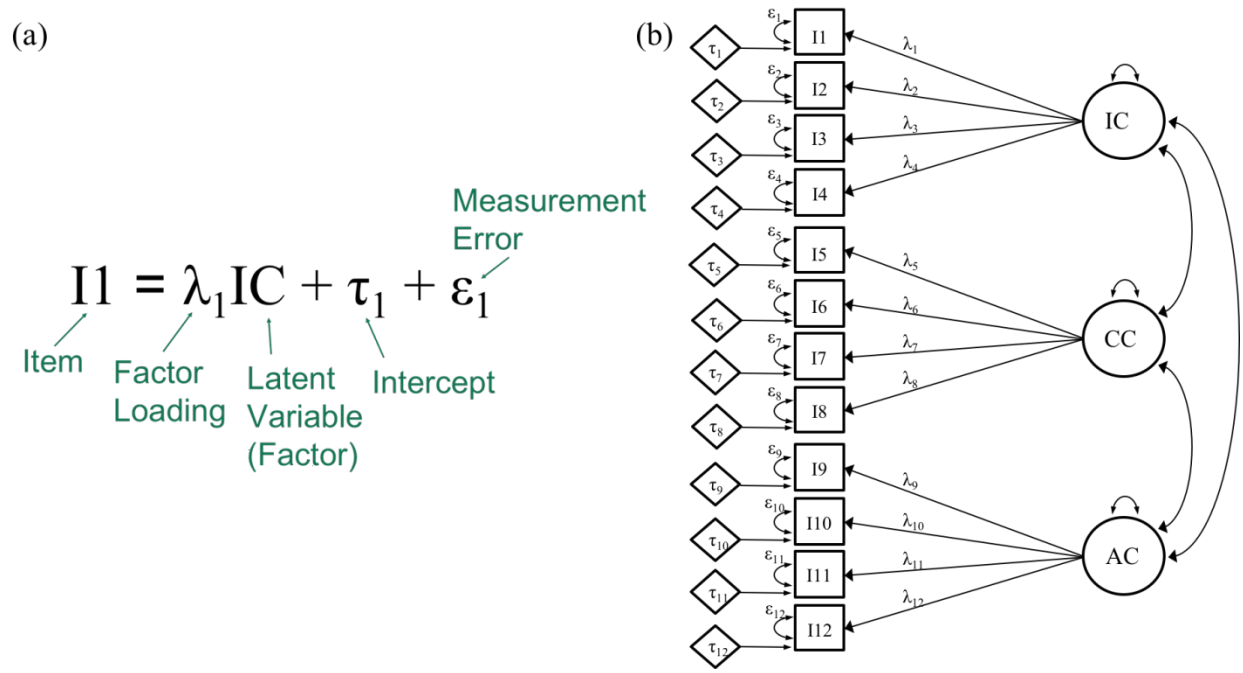


Figure 6. (a) Representation of equation components in CFA. Linear equation for I1 and the IC factor with notation and labels corresponding to CFA framework. (b) Factor model displaying the factor analysis notation of the relation between items and their corresponding factors.

Measurement invariance testing within a CFA framework investigates the extent to which the network of equations in a model is similar across group-level data. Therefore, each part of the equation (Eq. 1), for each item is tested for evidence of significant differences across groups, starting with the slopes (loadings), then the intercepts, followed by the measurement error variances. At each stage of measurement invariance testing, evaluation of overall data-model fit occurs.

Data-Model Fit and Fit Indices

The primary goal of measurement invariance testing is to examine how well the data collected fit a proposed model of relations among items and factors as described by a set of regression equations. Continuing with the example, we investigate the PRCQ data (by STEM and non-STEM groupings) for item associations based on the proposed (*a priori*) three-factor model for the PRCQ shown in Figure 6b. Mapping the data to this proposed model using a maximum likelihood estimator (the default in most software packages and the one that is appropriate for our simulated continuous data), fit indices are generated and are used to evaluate how well the data fit the model. Regardless of the software package used, it is good practice to review several kinds of fit indices that fall in each of these categories: comparative fit, absolute fit, and parsimony correction. The comparative fit indices evaluate the fit of a specified model solution in relation to a baseline model solution. Absolute fit indices assess how reasonable the model fit is based on the null hypothesis that the data fit the model perfectly. Finally, the parsimony correction indices are similar to the absolute fit but include a penalty for poor model parsimony (Brown, 2006).

With these fit index descriptions in mind, we present several suggested cutoff criteria for fit indices that were simulated by Hu and Bentler (1999) using a maximum likelihood estimator. Examples of comparative fit are: the Comparative Fit Index (CFI) and the Tucker-Lewis Index (TLI), both of which have a recommended cutoff of >0.90 as acceptable, but best if >0.95 (Hu and Bentler 1999). For the absolute fit category, the Chi-square (χ^2) test statistic and the standardized root-mean square residual (SRMR) indices can be considered. The χ^2 is a descriptive index utilized to evaluate how closely the data fit the model. However, this test is highly influenced by sample size, thus additional fit indices must be considered to evaluate appropriate data-model fit (Brown, 2006). Hence the SRMR is a valuable index to add in this category and its cutoff criteria is <0.08 as acceptable (Hu and Bentler 1999). Finally, for the parsimony correction, the root mean square of approximation (RMSEA) index can be evaluated with acceptable cutoff criteria of <0.06 (Hu and Bentler 1999). Though these recommended criteria are often considered as firm cutoffs, there are known situations where the strength of the factor loadings can confound interpretation of fit indices (McNeish *et al*., 2018). Therefore, it is up to the researcher to provide as much evidence as possible to support the acceptability of a proposed factor model. It is also important to note that for categorical data a different estimator should be used, thus model fit indices and cutoff criteria are different from the ones noted here for continuous data and the maximum likelihood estimator. A more thorough description of estimator, model fit indices, and their respective cutoffs for categorical data are provided in the ESI.

In the following section of this manuscript we present measurement invariance testing as the step-by-step evaluation of a series of nested models. Each step in the evaluation adds a constraint to test whether the groups being compared share a similar measurement model and if comparisons can be supported. Therefore, in addition to evaluating the data-model fit at each step of measurement invariance testing, we also calculate and evaluate the change in data-model fit between nested models. Cheung and Rensvold (1999, 2002) as well as Chen (2007) conducted a series of simulation studies with continuous data to investigate data-model fit criteria, in particular the change in data-model fit at each step of measurement invariance testing. Cheung and Rensvold (2002) focus solely on evaluating the change in Chi-square ($\Delta\chi^2$) between nested

models, looking for a nonsignificant value. More recent work finds this practice acceptable (Mueller and Hancock, 2019), as the idea of measurement invariance testing is to find no evidence of significant difference between the models, which provides support for group comparisons. Other researchers, such as Chen (2007) have investigated the change in other fit indices as well, to ensure that there are various indicators that provide further evidence that no significant difference between nested models is observed. Chen (2007) offers a range of values that, based on the simulation studies conducted, offer reasonable cutoff values for the fit indices we have introduced earlier in this section. These values vary by level of invariance being evaluated and therefore will be presented within the appropriate testing step below. However, simulation studies have called into question the exact cutoffs and fit indices to use in the context of invariance testing (Kang *et al*., 2016) so again the researcher must decide what evidence to present to justify interpretation of models.

Steps of Measurement Invariance Testing

In 1997 Widaman and Reise described 4 steps of measurement invariance testing: configural, metric (weak), scalar (strong), and residual (strict, also known as conservative). In this report we focus on this 4-step method, although there are other methods that utilize additional steps when investigating whether comparisons are supported between groups (for examples see Jöreskog, 1971; Vanderberg and Lance, 2000).

Step 0: Establishing Baseline Model

A preliminary step before conducting measurement invariance testing is to conduct a separate CFA for each group dataset that will be compared. In this step, the CFA is used to investigate that each group's response patterns align with the proposed model to an acceptable level (Gregorich, 2006). The acceptability of the fit between each dataset (i.e., STEM and non-STEM groupings) and the model (Figure 6) is checked using the fit indices noted earlier. If the data-model fit for either group's data is deemed unacceptable at this stage, measurement invariance testing is not appropriate and comparisons between the groups would not be supported. At this point, the next step would be to conduct an investigation of the reasons for failing to achieve acceptable data-model fit. However, if the data-model fit reached acceptable criteria for each group, then beginning the measurement invariance testing steps is appropriate.

Step 1: Configural Invariance

Once the independent CFAs for each group are found to have acceptable data-model fit, the first step of measurement invariance testing can begin. In this step, the same model is estimated concurrently for each group, allowing all model parameters to be freely estimated (Gregorich, 2006; Sass, 2011; Putnick and Bornstein, 2016). The point of this unconstrained model is two-fold: 1) to investigate whether items associate with each other in similar ways in all groups (i.e., items belonging to the same factor correlate more highly with each other than to other items); and 2) to establish a baseline of data-model fit, ensuring that subsequent comparisons are conducted utilizing the same network of equations for both groups. This baseline model is called the *configural model*, as it verifies that the general structure (or configuration) of items and factors is similar across groups. Configural invariance is achieved when this model has acceptable data-model fit values (Hu and Bentler, 1999).

The models in Figure 7 represent the configural model, for our three-factor PRCQ instrument, for two groups. For discussion purposes, the model parameters for STEM majors (group 1) are labeled with numeric subscripts and those for non-STEM majors (group 2) are labeled with alphabetical subscripts. Take for example the relation between the first factor, IC, and the first item, I1. This relation is symbolized as λ_1 for group 1 and λ_a for group 2. In the configural model, these two relations are free to take on whichever value provides the optimal solution to the system of regression equations.

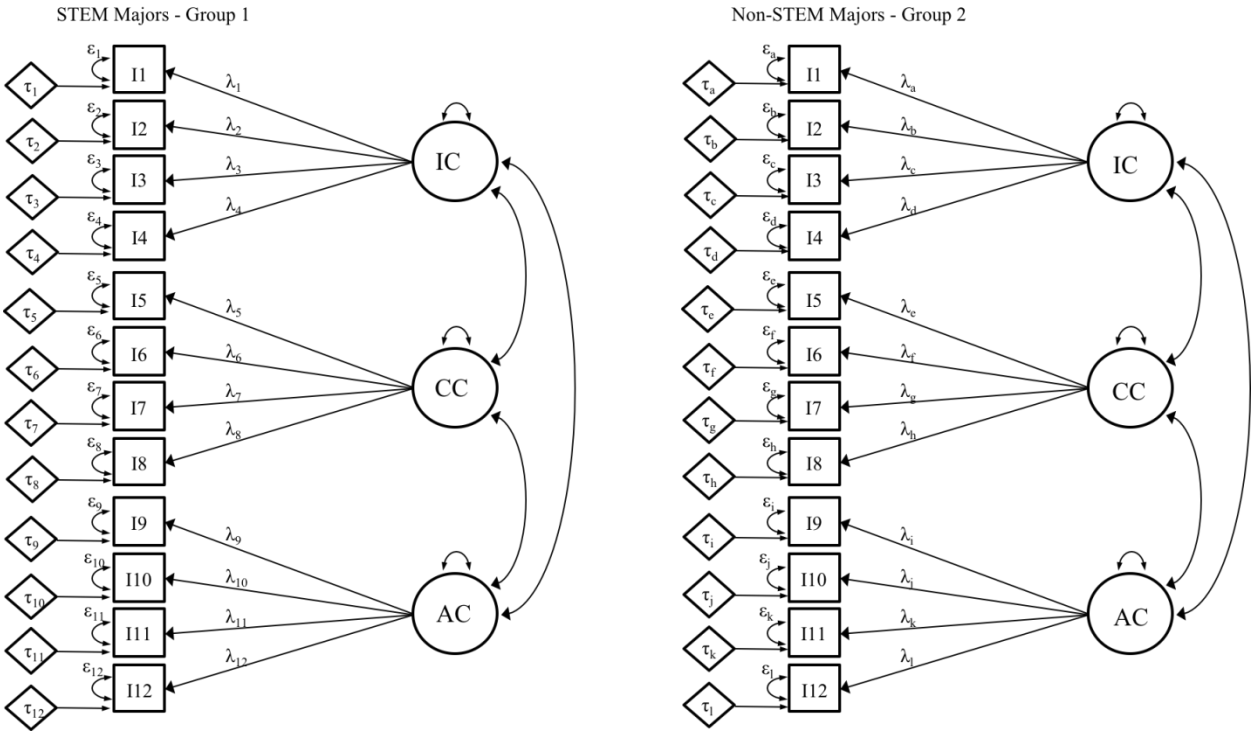


Figure 7. Configural invariance model where all parameters are freely estimated for two groups (STEM and non-STEM majors).

If the configural model fails to reach acceptable levels of fit, the result suggests that the factors are not associated with the same items for both groups (Gregorich, 2006; Putnick and Bornstein, 2016). Therefore, one can question whether the constructs being measured have the same meaning for these groups (Bornstein, 1995; Putnick and Bornstein, 2016). With this outcome, no further invariance testing is advised. However, we encourage researchers to conduct further investigation to find the source of noninvariance between the groups. Modes of investigation could be quantitative in nature, such as inspection of covariance or correlation matrices similar to the visuals we provided earlier (Figures 1- 4). Investigation could also be qualitative in nature, for example conducting cognitive interviews (Willis, 1999) with respondents from both groups to explore the constructs being measured and find the root of the differences between the two groups. These practices can help to ascertain any fundamental differences in construct meaning for different groups, which can provide insight into their lived experiences and interpretation of the construct of study (Komperda, Hosbein and Barbera, 2018).

Step 2: Metric Invariance (Weak)

If a configural model (Figure 7) is observed to have acceptable data-model fit, the next level of establishing equality between the group-level data can be conducted. This step involves applying the first constraint to the baseline model equations, which establish the linear relationship between items (e.g., I1) and factors (e.g., IC). In the *metric model* (Figure 8), also called the *weak invariance model* (Meredith, 1993), the constraint of equal unstandardized slopes, or factor loadings (λ), is applied (Gregorich, 2006; Sass, 2011; Putnick and Bornstein, 2016; see Figure 8 where loading subscripts match across groups). That is to say that for STEM majors, the factor loadings are freely estimated, but for non-STEM majors, the loadings are set to be equal to the loadings for STEM majors. At this level of invariance testing, we are exploring whether the strength of associations between the items and the latent variables are similar across groups (Byrne, Shavelson and Muthén, 1989; Gregorich, 2006). To achieve metric invariance, first the fit statistics of the metric model (Figure 8) are evaluated (Hu and Bentler, 1999), and then they are compared to those of the configural model (Figure 7). No evidence of significant difference should be observed between the configural and metric models. To evaluate the comparison between configural and metric models, the change in fit indices between levels is established utilizing the guidelines noted earlier. It is important to note that evaluating model fit is pertinent; however, evaluating the change between the models is essential to establishing invariance between groups. Establishing metric invariance implies that the meaning of the factor (in terms of relative weight of items) is similar across groups (Gregorich, 2006). However, this evidence is not enough to make comparisons between groups. At the very least, another level of constraint is needed before group comparisons can be made, as will be summarized in subsequent steps.

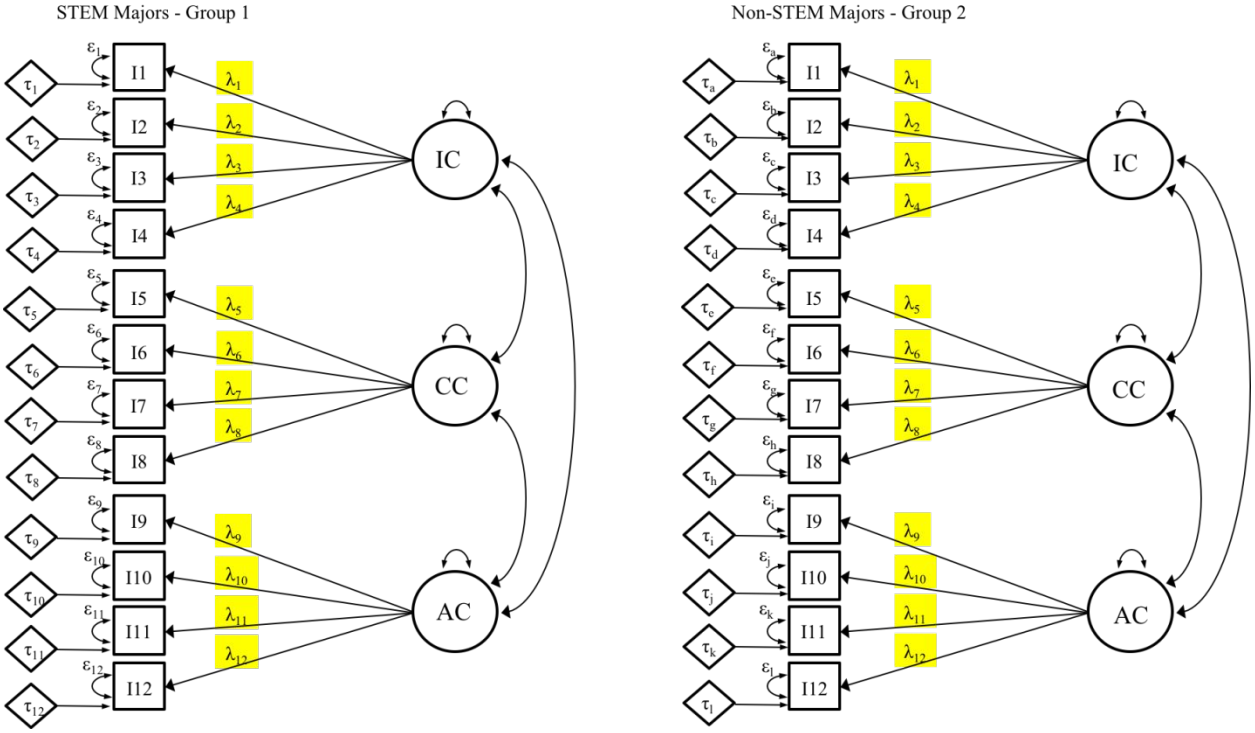


Figure 8. Metric model where factor loadings (highlighted) are constrained to be equal for both groups. All other parameters (e.g. intercepts and error variances) are freely estimated.

Failure to reach metric invariance suggests that the strength of association between items and the factor to which they belong are different between the groups. The strength of item association with the factor provides meaning to the factor from the perspective of the respondents (Gregorich, 2006). Therefore, if the item-factor associations are significantly different across groups, then the meaning of the underlying factor is different between groups, or the factor loadings are biased (Gregorich, 2006). Generally, when metric invariance is not achieved, there are one or more items with poor loadings for one of the groups compared to the other group. At this juncture, investigation of the item loadings or modification indices generated by the software can provide meaningful insight about the different ways that respondents may associate items to the underlying construct. After evaluation, researchers may choose to release the constraint of equal loadings for the problematic item(s) and run the model again for partial measurement invariance (Byrne, Shavelson and Muthén, 1989; Putnick and Bornstein, 2016). If this release of constraints is undertaken, comparisons between groups are cautioned, particularly for the constructs that involve the problematic items. These items might be the subjects of further investigation as to the alignment between items and underlying constructs for the groups of interest.

Step 3: Scalar Invariance (Strong)

Once metric invariance is established (i.e., no evidence of significant difference is found between the metric and configural models), the next constraint can be applied. The *scalar model* (Figure 9), also called the *strong (factorial) invariance model* (Meredith, 1993), consists of incorporating unstandardized equal intercepts, in addition to equal loadings, across groups in the model (Gregorich, 2006; Sass, 2011; Putnick and Bornstein, 2016). With this addition, the intercepts (τ) are freely estimated for STEM majors, but for non-STEM majors they are set to be equal to the intercepts for STEM majors (see Figure 9). The purpose of this model is to establish evidence of unbiased estimated factor mean differences between groups (Gregorich, 2006), which implies that factor means encompass all mean differences in the shared variance of the items (Putnick and Bornstein, 2016). Factor means are unbiased because the error terms (ε) are not part of them. This is not true for observed item and observed scale means as they are calculated from the observed item scores that include the associated error terms (Putnick and Bornstein, 2016).

Just as with the metric model, first the scalar data-model fit is evaluated (Hu and Bentler, 1999) and then the fit comparison between, now, the metric (Figure 8) and scalar (Figure 9) models utilizing the appropriate values noted earlier. We reiterate that evaluating data-model fit is an important step of measurement invariance; however, essential to providing sufficient evidence for score comparisons is the change in fit statistics from one model to the next.

Once scalar invariance is achieved, the researcher has established evidence to support the comparison of *factor means* between groups. This evidence helps to rule out that any observed differences arise from variations caused by systematic higher or lower item responses (Gregorich, 2006; Sass, 2011; Putnick and Bornstein, 2016) due to issues like cultural norms.

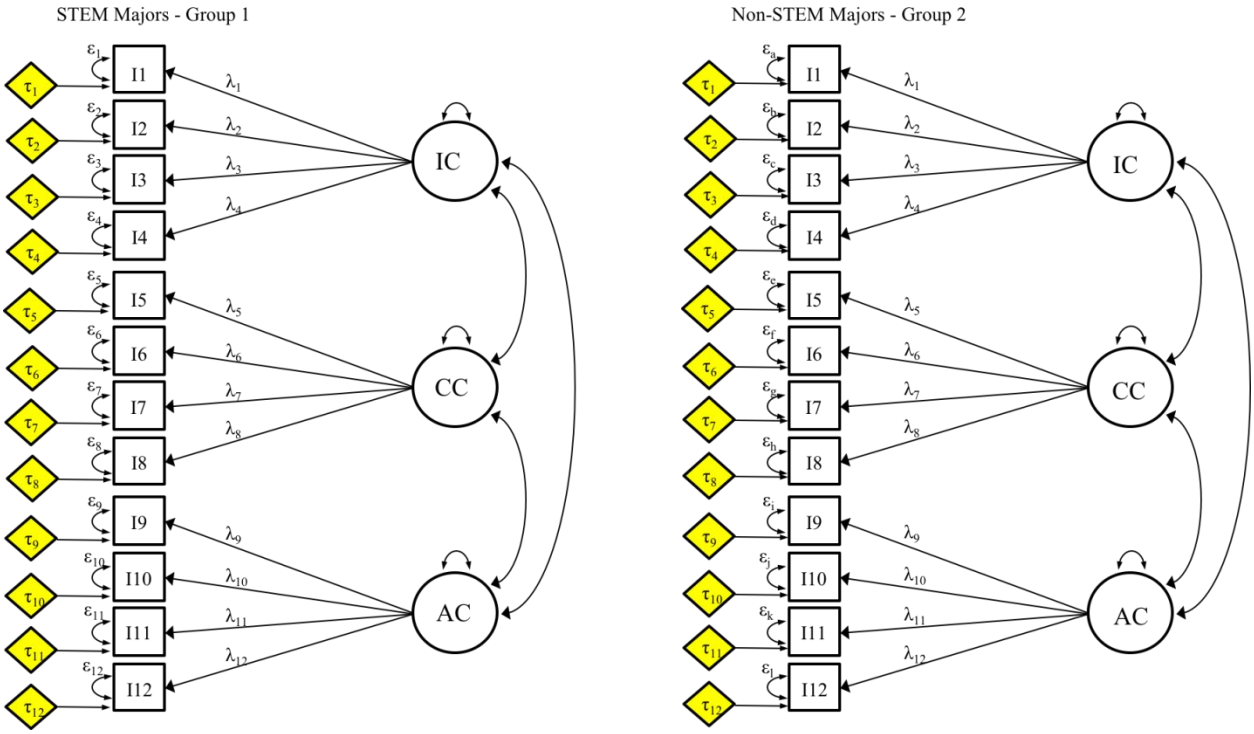


Figure 9. Scalar model where factor loadings and intercepts (highlighted) are constrained to be equal for both groups. All other parameters, including error variances are freely estimated.

If the scalar model provides results that are significantly different from the metric model, then scalar invariance has not been achieved and factor mean comparisons between groups are not supported. However, investigation as to the source of mismatch can be conducted. As demonstrated earlier, visualizing the data can be helpful at this juncture. Figure 5 shows item intercepts displayed as boxplots. Although one can choose to visualize data in various ways, Figure 5 visually suggests that the intercept for I3 (STEM majors) might be different than the intercept of the same item for the non-STEM majors. As I3 belongs to the IC factor, interpreting the IC factor mean comparisons between groups can be more difficult given this limitation. However, investigation as to the reason for the mismatch between groups is warranted. As previously mentioned, differences in item intercepts can be caused by diverging cultural norms that cause higher or lower item responses in diverse groups (Gregorich, 2006), thus investigating the source of the difference is encouraged. An example of this phenomenon that could cause systematic higher or lower responses is acquiescence bias. For example, one group might not utilize the entire response scale range, rather the response distribution is skewed to either end of the scale or narrowly in the middle.

In this situation, researchers may choose to release the constraint of equal intercepts for I3 only and evaluate the scalar model again. If releasing the constraint for I3 results in scalar model fit that is not significantly different from the metric model, then scalar invariance is established *with limitations*, sometimes described in terms of partial invariance (Putnick and Bornstein, 2016; Fischer and Karl, 2019). However, if an item loading was not held constant between groups in a previous step of invariance testing then the intercepts must also not be held constant as there is no reason to believe items with two different slopes would be expected to

have the same intercepts. There is some evidence that with partial invariance of intercepts comparison of factor means may provide acceptable results (Steinmetz, 2013).

An important distinction at this juncture is that factor means are obtained from the model, not from summing or taking the average of the observed item response values. Factor means are not a 'set' number, rather they are a comparison of latent (unobserved) means between two (or more) groups, where one group serves as the reference, taking the value of zero, and the other group or groups is/are compared to the reference. An effect size of the comparison can also be calculated (Hancock, 2001; Bunce *et al.,* 2017). Although this way of making comparisons is not frequently used in CER, the application of this practice is useful. We encourage researchers to work with factor means more often for two main reasons: 1) As explained earlier, factor means are estimated from the model, capture all mean differences in the shared variance of the items in the factor, and are free from error terms (Putnick and Bornstein, 2016). This cannot be said for observed scale scores, meaning composite scores taken directly as an average or sum of the observed variables (i.e., items), since these scores must include the error terms and do not take into account the strength of the association between items and factors. 2) In order to compare observed scale scores, the conservative invariance test, described in the following section, must be achieved. Meaning, it is harder to provide sufficient evidence for observed scale score comparison between groups than it is to compare factor means. Thus, we encourage researchers to utilize factor means as an effective tool for group comparisons as these values are void of error terms and will lead to more accurate interpretations and more meaningful inferences.

Step 4: Conservative Invariance (Strict)

Once scalar invariance is achieved, comparison of factor means between the groups is possible. However, if researchers desire to compare the observed scale scores of each factor; meaning composite scores taken directly as an average or sum of the observed variables (i.e., items), it is advisable to conduct a *conservative* or *strict* (Meredith, 1993) invariance test first (Gregorich, 2006; Sass, 2011). The conservative test checks the additional condition that measurement error variances are similar across groups. This is done in the same fashion as the prior models, with the final addition being that the STEM majors' error variances (ε) are freely estimated and non-STEM majors' error variances (ε) are constrained to be equal to those of STEM majors (see Figure 10). At this point, all loadings, intercepts and error variances are fixed to be equal between the groups to be compared. To establish strict invariance, the data-model fit statistics are first evaluated and then compared between the strict (Figure 10) and scalar (Figure 9) models and no evidence of significant difference should be found. If strict invariance is established, enough evidence is gathered to warrant observed scale score comparisons between groups (Gregorich, 2006; Sass, 2011). This type of comparison is what most researchers are accustomed to investigating; however, it is important to note that these comparisons require evidence of meeting this highest level of invariance testing. Failure to achieve strict invariance means that observed scale comparisons are not supported. Thus, researchers may investigate scalar invariance (i.e., Step 3) to compare factor scores instead.

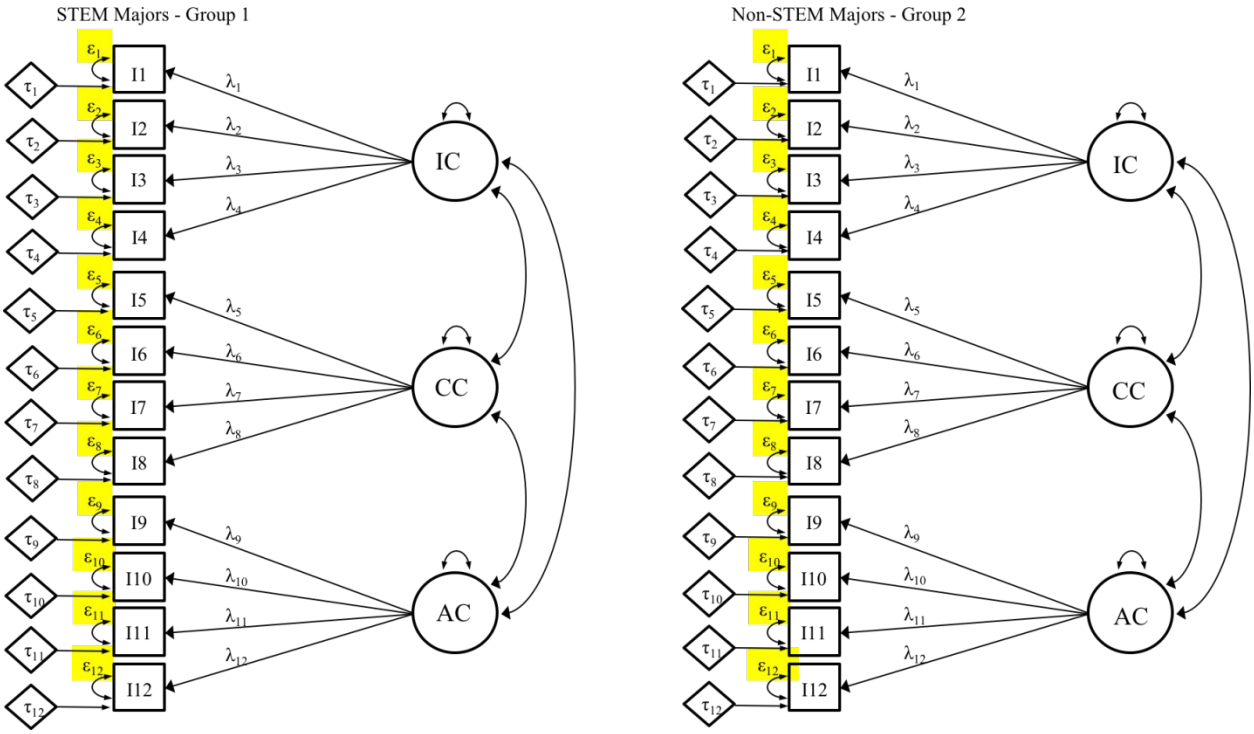


Figure 10. Conservative (strict) invariance where loadings, intercepts, and error variances are constrained to be equal for both groups.

Based on the four steps described previously, we provide a summary table (Table 1) for readers to reference as they conduct measurement invariance testing in their own studies. This table, while not comprehensive, provides the basic model characteristics, the evidence established, appropriate claims, and supported group comparisons that can be made at each level of invariance testing. This table can also prove useful as reviewers and journal editors review quantitative studies that can benefit from this method to support comparisons between groups or across time.

Table 1. Summary of Claims and Evidence Established at Each Stage of Measurement Invariance Testing - Guide for Researchers, Practitioners, and Reviewers.

	Configural	Metric (Weak)	Scalar (Strong)	Conservative (Strict)
Model characteristics	all parameters freely estimated in all groups, no constraints	factor loadings constrained to be the same for all groups	factor loadings and item intercepts constrained to be the same for all groups	factor loadings, item intercepts, and error variances constrained to be the same for all groups
Evidence established	same number of factors, items associated with the same specific factor for all groups	evidence in configural **plus** same strength of association between factors and corresponding items for all groups	evidence in configural and metric **plus** same item intercepts for all groups	evidence in configural, metric, and scalar **plus** same item error variances for all groups
Appropriate claims	items are associated with each other and the underlying factors in similar ways	claims from configural **plus** meaning of the factor (in terms of relative weight of items) is similar across groups	claims from configural and metric **plus** no systematic response biases; differences in factor means are due to a true difference in groups	claims from configural, metric, and scalar **plus** no systematic response biases or difference in error between groups; differences in item and scale means are due to a true difference in groups
Supported comparisons between groups	none	none	factor mean scores (from the model)	observed scale scores

Measurement Invariance Testing Example with Simulated Data

To illustrate the steps of utilizing measurement invariance testing for determining if, and to what degree, group comparisons can be made, we use the simulated dataset that generated Figures 4 and 5 to work through an example. The data was simulated to be continuous, therefore the maximum likelihood estimator was used for each model. For each step in the process, the data-model fit results as well as the fit comparisons between models are displayed in Table 2. It is important to note that while fit indices for each model will be calculated and tabulated by the software being used (i.e., R or Mplus, etc), the change values between models have to be manually calculated with a simple subtraction, with the exception of the *p*-value associated with the $\Delta\chi^2$, which must be retrieved from a χ^2 table that contains degrees of freedom.

At the baseline (Step 0) and configural (Step 1) levels, only the overall data-model fit is investigated. In our PRCQ example, the data at these levels was simulated with essentially perfect data-model fit as noted in Table 2. Perfect fit at these levels is unlikely to happen in a real study; thus, expecting a less-than-perfect fit is reasonable. Therefore, evaluating the data-model fit should follow acceptable guidelines, such as those by Hu and Bentler (1999) used here, or others as appropriate based on the data type. As each of our independent baseline models showed acceptable data-model fit and then the combined configural model showed good data-model fit, we can proceed to the next step of invariance testing.

The metric model data (Step 2) exhibits acceptable data-model fit (see Table 2). Beginning with these metric level indices, we not only evaluate the data-model fit but also compare the fit obtained with the metric model to that of the configural model. First, we evaluate the $\Delta\chi^2$ (Cheung and Rensvold, 2002; Mueller and Hancock, 2019) which is a non-significant value, thus providing proof that there is no evidence of significant difference between the models. Then, following the suggestions of Chen (2007), our calculated values of ΔCFI = 0.000, ΔSRMR = 0.001, and ΔRMSEA = 0.003 are within the acceptable change cutoff levels: ΔCFI (< 0.01), ΔSRMR (< 0.03), and ΔRMSEA (< 0.015) to establish metric invariance (Chen, 2007). The comparison between configural and metric models shows that there is no evidence of significant change between these two models, thus metric invariance is achieved based on the comparison and we are warranted in moving to the next step of invariance testing.

For evaluating if scalar invariance is achieved (Step 3), a similar analysis pattern is followed. First, we evaluate the data-model fit. At this point, we observe that the fit indices for the scalar model are no longer within the acceptable ranges (see Table 2). This result is problematic because it is an indication that scalar invariance does not hold for the groups. Further evidence is found when we compare the change in fit indices between the metric and scalar models. Here we observe that our value of $\Delta\chi^2$ is significant, and the values for ΔCFI, ΔSRMR, and ΔRMSEA are also not within the recommended fit index cutoffs: ΔCFI (< 0.01), ΔSRMR (< 0.01), and ΔRMSEA (< 0.015) for scalar invariance (Chen, 2007). These additional results confirm that scalar invariance is not reached for these data. As the model at this level is not supported, we do not go on to evaluate the next highest level of invariance (i.e., the strict invariance model at Step 4), as we do not have a supported scalar model to compare it to. However, if the scalar model held and we desired to move on to test for strict invariance, the same guidelines and fit index cutoffs would be used as for scalar invariance (Chen, 2007).

In this simulated data example with the PRCQ, our analysis provided evidence for metric invariance at Step 2 but not for scalar invariance at Step 3. Therefore, these results imply that factor mean comparisons between STEM and non-STEM majors are not supported and should not be performed. Investigating the source of the misfit in the scalar model is warranted. Based on our previous discussion, we know that the I3 intercept is higher for the STEM majors compared to non-STEM majors (see Figures 4 and 5). At this point, we may choose to qualitatively investigate the difference between these groups for I3. Alternatively, we may choose to release this item's intercept constraint (i.e., allowing the I3 intercept for each group to be freely estimated) and run the scalar model again. If the data-model fit and model comparisons indicate acceptable levels with this modification, *partial* scalar invariance would be achieved. At this point, we would have limited support for factor mean comparisons. However, we would not be able to make any significant claims, particularly for the IC factor, due to the limitation for I3. Based on this limitation, reflection on the consequences of making factor mean comparisons between these groups and the validity of inferences drawn from these comparisons is crucial. Finally, as we were not able to evaluate for scalar invariance, we have no basis for comparing the observed scale scores of the STEM and non-STEM majors using the PRCQ.

Table 2. Measurement Invariance Testing for the PRCQ Instrument Comparing STEM Majors and Non-STEM Majors with Simulated Data for Illustration

Step	Testing level	$\chi2$	*df*	*p*-value	CFI	SRMR	RMSEA	$\Delta\chi2$	Δdf	*p*-value	ΔCFI	ΔSRMR	ΔRMSEA
0	STEM majors Baseline	65	51	0.084	0.998	0.021	0.017	-	-	-	-	-	-
0	Non-STEM majors Baseline	52	51	0.437	1.000	0.016	0.004	-	-	-	-	-	-
1	Configural	117	102	0.142	0.999	0.018	0.012	-	-	-	-	-	-
2	Metric	120	111	0.245	0.999	0.019	0.009	3	9	0.231	0.000	0.001	0.003
3	Scalar	2268	120	< 0.001	0.820	0.191	0.134	2148	9	< 0.001	0.179	0.172	0.125

Note. STEM majors $n = 1000$. Non-STEM majors $n = 1000$. Simulated data was used and altered at the scalar level (intercepts) for illustrative purposes; fit indices are from R.

As we have described throughout this manuscript, and shown through the example here, measurement invariance testing provides researchers and practitioners with statistical evidence to support (or in this case, refute) comparisons between the groups evaluated (Sass, 2011). Once it has been established that both groups view the items on an instrument in similar ways (i.e., by establishing a certain level of measurement invariance), the interpretation of results becomes validated. Utilizing measurement invariance testing provides support for meaningful inferences between populations, taking into account response patterns that may arise from a group's background or experiences (Wicherts, Dolan and Hessen, 2005). Furthermore, providing evidence that the data from an assessment instrument does not have validity threats against a comparison group, such as URMs (Gillborn *et al.,* 2017), provides more confidence in the results obtained and may provide increased support for claims of social inclusion for these groups.

Limitations

While we encourage all researchers and practitioners to utilize measurement invariance testing prior to conducting comparisons between groups, we acknowledge there are limitations which may not allow the use of this method. One of these limitations is the sample size required to conduct these model-based tests. Similar to factor analysis techniques, measurement invariance testing requires a large sample size. Although there are no specific rules about the sample size required, some work indicates that sample size requirements can be calculated given the number and value of the parameters being estimated (Wolf *et al.,* 2013; Mueller and Hancock, 2019). However, we encourage researchers to continue investigating newer methods to determine appropriate sample size that may be more suitable for this technique that is specific to the model parameters, the type of data being analyzed, and other characteristics of their study (Wolf *et al.,* 2013). Therefore, when conducting research and comparisons between groups with small samples, the technique presented in this work is not appropriate. Thus, we encourage researchers, practitioners, reviewers and journal editors to consider other methods of reflexivity such as response process evidence and/or content review by culturally-aware experts.

Additionally, researchers should be aware that the fit index cutoffs we have presented in this manuscript both for evaluating data-model fit and for change in model fit indices are suggested values based on simulation studies. While these guidelines are generally accepted within the field of measurement, this is an area of active investigation and these guidelines could evolve in coming years. As we encourage researchers to follow these guidelines, we also encourage a thoughtful evaluation of the data, model, and data-model fit where the suggested guidelines may not apply (Kang *et al*., 2016; McNeish *et al*., 2018).

Another limitation of measurement invariance testing is that this technique alone does not inform the exact ways in which groups differ in item and factor interpretation. Although this technique can point to the problematic items and factors that are dissimilar between groups, it cannot provide reasoning for the different meaning of items or factors between groups. This information is best investigated using qualitative methods that can inform the perspective and interpretation from a respondent's point of view.

Finally, as with all statistical inferences, the measurement invariance testing process is built upon a series of assumptions. Without clearly identifying and acknowledging these assumptions, there is little support for the conclusions drawn from invariance testing. Due to the limited focus of this manuscript, only a few of the underlying assumptions for invariance testing were briefly discussed (i.e., theoretical support for the model being tested, quality of data being fit to the model, and acceptability of partial invariance at the metric and scalar stages). However, other assumptions are described more fully in the ESI and other resources (Bontempo and Hofer, 2007; Hancock *et al.*, 2009; Putnick and Bornstein, 2016; Fischer and Karl, 2019).

Discussion

CER is moving in the direction of greater interest in the differential impacts and outcomes of diverse populations (Rath *et al.,* 2012; Fink *et al.,* 2018; Stanich *et al.,* 2018; Shortlidge *et al.,* 2019). However, efforts to increase diversity by enrolling more URM students are not sustainable unless paired with efforts to increase social inclusion and social justice (O'Shea *et al.,* 2016; Puritty *et al.,* 2017). In an effort to 're-imagine' quantitative approaches to

better serve social justice initiatives (García, López and Vélez, 2017) and raise the standards for investigating these issues at different intersections of identity and background (e.g., race and gender; race and math preparation, etc.), we have presented a statistical method which investigates potential validity threats that could arise when analyzing assessment instrument data. Particular focus has been given at each stage of the analysis to explain some issues that could be evidenced if a given model fails to reach acceptable data-model fit criteria. We have included a few examples that could provide readers some ideas to begin their investigation when measurement invariance is not established at a particular level. Suggestions for circumventing some of these difficulties, such as releasing individual item parameters, have also been presented along with their implications. A summary of each stage of testing, along with the supported claims and evidence established is provided in Table 1.

Many recent studies in CER have taken the first step toward raising the research standards by including variables such as gender, race, etc. and appropriate intersections in their studies (Rath *et al.,* 2012; Fink *et al.,* 2018; Stanich *et al.,* 2018; Shortlidge *et al.,* 2019). However, the next step of investigating the validity of the group comparisons was lacking. Therefore, we encourage researchers to investigate their own data, even the data that has already been published, and consider whether the inferences made were valid for the populations being compared. One recent example of this practice is the study conducted by Rocabado and colleagues (2019), which explored data from a study done in 2016 by Mooring and colleagues who conducted an evaluation of the attitude impact of an organic chemistry flipped classroom compared to a traditional classroom. The researchers found that the flipped classroom showed significant attitude gains when compared to the traditional classroom (Mooring *et al.,* 2016). Rocabado and colleagues (2019), not only investigated whether the original comparison was supported, but also studied whether the attitude gains observed extended to the Black female students in the original sample by utilizing measurement invariance testing to support the investigation and comparisons.

Measurement invariance testing provides opportunities to investigate levels of differences that could arise any time group comparisons are to be made. The 4-Step method presented in this primer is not limited to group comparisons by gender, race, or ethnicity only, it includes groups such as those used in this manuscript (i.e., STEM and non-STEM majors) and to same-group analyses in longitudinal comparisons (e.g., pre-post gain). Regardless of how the groups are defined, at the configural model level (Step 1), an acceptable data-model fit suggests that the groups utilize the same network of equations and the basic measurement model (e.g., number of factors present). At this stage, the claim can be made that item associations are similar between groups, as demonstrated by Figure 2. The configural model provides a lens to observe these item associations when the data is disaggregated by the defined groups. Item correlations might not be similar for all groups and therefore, the configural model might not reach acceptable levels of data-model fit, suggesting group-level differences in the constructs being measured. If this level cannot be achieved, comparisons between groups are not fair due to the difference in constructs. This is an important step in measurement invariance testing, as it provides a strong foundational model on which to base the subsequent tests.

The metric model (Step 2) investigates the strength of the association between factors and their corresponding items (Sass, 2011). The strength of these relations indicates the meaning of

the factor (Gregorich, 2006). Therefore, when the metric model fails, it is evidence of differences in factor meaning between the groups, which provides grounds for further investigation. These differences are observed when the entire pattern of item loadings differs between groups. As this result does not indicate why the groups differ in meaning, a thorough investigation of data from items and constructs should be reviewed for validity evidence including aspects of content validity, response process validity, and construct validity, keeping in mind the various groups that could be in the target population. Metric non-invariance may also arise when one or more item loadings on a factor differ greatly between groups (see Figure 3c), indicating that one group does not associate the item(s) with the construct being measured, while the other group does. For example, in the *fictitious* Applications of Chemistry (AC) scale, a problematic item might ask about the field of Materials Science. As this field is interdisciplinary between Engineering, Physics, and Chemistry, it is likely that STEM students would have been exposed to examples from the field across many courses. However, non-STEM majors may have never been exposed to the ideas and examples of Material Science and the role Chemistry plays. Therefore, when comparing a group of STEM majors, who are more likely to have been exposed to Materials Science, to a group of non-STEM majors, it is possible that this item functions differently between the groups. The non-STEM majors might not view Material Science as being an application on the AC scale because they have not been introduced to this field and its interconnections. Therefore, when an item cannot be explained by the underlying construct for one group, the meaning of the construct is different between the groups.

The scalar model (Step 3) considers whether item averages within the measurement model are similar across groups. As shown in Figure 4, item averages may look similar when combined; however, when disaggregated into groups, item means could be different (Figure 5) leading to the scalar model not reaching acceptable levels of fit. These differences could arise due to acquiescence biases that affect one group and not the other due to cultural norms not shared between groups (Gregorich, 2006). In the *fictitious* Connectedness of Chemistry (CC) scale, a problematic item might ask about the degree to which chemistry is connected to a specific issue of global warming, say CO_2 emission. One can think that STEM majors might see stronger ties between the issue and chemistry and therefore score higher on this item than a group of non-STEM majors that may not have been exposed to the idea of light-matter interactions. Therefore, if all the STEM majors score this item high (i.e., a 4 or 5 on a 5-point scale) because they have learned about this phenomenon, then the scale is biased for this item between the two groups in this context. If scalar invariance is not achieved, comparisons between groups beyond the metric model level are not warranted. On the other hand, if scalar invariance is reached, estimated factor mean scores can be computed and compared between groups with evidence that differences between groups are not artifacts of the instrument and construct meaning is similar across the groups. However, if a researcher's goal is to compare observed factor scores (e.g., observed item averages), evidence of conservative invariance (Step 4), in which error variances are constrained to be equal between groups, is required (Sass, 2011).

While conducting measurement invariance testing, each stage provides safeguards and reflexivity (Gillborn *et al.,* 2017) about the groups being compared, rendering this quantitative approach suitable for investigating the differential impacts and outcomes of diverse populations and advancing social justice and equity in CER at the institutional level. We encourage all researchers and practitioners not only to investigate the impact of variables such as race/ethnicity

and appropriate intersections (e.g., gender status, language status, socioeconomic status) more often in their research and in their classrooms, but also to employ techniques such as measurement invariance testing in order to safeguard against disguising racism and other social injustices and systemic biases when making comparisons between groups (Gillborn *et al.,* 2017; García, López and Vélez, 2018).

Recommendations and Implications

Measurement invariance testing provides evidence to support or refute quantitative data any time group comparisons are to be made. Although qualitative methodologies are used more often to investigate individuals' and groups' lived experiences, utilizing quantitative methods with reflexivity and safeguards against racial and other biases (Gillborn *et al.,* 2017; García, López and Vélez, 2018) can enhance research and teaching that aims at studying pedagogies and interventions that benefit URMs in chemistry. This quantitative method is not limited to group comparisons by gender, race, or ethnicity. It includes groups such as those defined by academic major, socioeconomic status, transfer status, or other meaningful categories and also extends to same-group analyses in longitudinal comparisons (e.g., pre-post gain). To make the endeavor of utilizing measurement invariance testing as easy and accessible as possible, we have provided code and ample explanation for two common software programs (R and Mplus) in the ESI. Although we provided code for these programs, there are a variety of other programs available that support this technique such as SAS, LISREL, EQS, or the AMOS add-in for SPSS. A helpful comparison of software for structural equation modeling with multiple groups can be found in Narayanan (2012).

For Researchers and Reviewers

Measurement invariance testing is a technique that we encourage all researchers to use when analyzing assessment instrument data for the purpose of group comparison in their studies. Identifying potential validity threats will greatly enhance the interpretation of the results obtained and claims made, as well as further the answer to the call for increased diversity and social inclusion. At each specific stage of measurement invariance testing, certain model claims can be supported or refuted, which either provide evidence for group comparison (see Table 1) or inform the subsequent steps to take in the research. Each of the measurement invariance steps is an opportunity to safeguard against observed and unobserved differences between groups that may be artifacts of the assessment instrument. As researchers, it is our duty to ensure that we present results that have the potential of being transformative; thus, working to minimize artifacts of measurement bias in our analyses is imperative to further the field of CER in more inclusive ways.

Likewise, when reviewing articles for publication, reviewers have the responsibility to ensure that the analyses conducted are held to high standards and that the results and implications are supported by sufficient evidence. In this work, we have highlighted the importance of conducting measurement invariance testing when researchers and practitioners utilize assessment instruments of latent traits on which groups will be compared. The results of these comparisons can have important implications and consequences in CER as the field moves toward greater diversity and social inclusion. Thus, these comparisons have to be made responsibly to properly address the consequential validity of the inferences drawn from studies where group or longitudinal comparisons are made. Particularly, we advocate for safeguards and

reflexivity in research methodology that aims to challenge the idea of neutral and objective research in an effort to work toward the abolition of social inequities (Solórzano, 1997; Yosso, 2005). Therefore, we urge reviewers and journal editors to check the conditions necessary for the comparison of outcomes by group. First, ensuring that researchers provide reason to believe it is valuable to compare the noted groups (i.e., the comparisons are not simply because the demographic data exists) on the variable of interest. Second, that there is reason to believe the construct being compared can be measured appropriately for all groups through establishing the relevant level of measurement invariance. We have shown how measurement invariance testing can provide reflexivity and ample opportunity to check for differences in measurement for groups in studies. Thus, we encourage the use of this method whenever possible.

Often, the comparisons made between groups will be done at the observed scale score level. If this is the ultimate goal of a study, then the researchers and reviewers should be aware that observed score comparisons require meeting strict invariance (the most conservative level of invariance) across all groups. If this strict invariance model provides acceptable data-model fit, then researchers and reviewers have evidence that observed scale scores can be compared between groups. Within this primer on measurement invariance testing, we laid out a step-by-step method, working up to establishing strict invariance. However, it is beneficial to mention that if only the strict invariance test is conducted, the investigation at each stage of measurement invariance testing is not provided and the change in data-model fit from one level to the next is not produced. Although valuable step-by-step information is not obtained when choosing to run only the desired test, this practice is sound. However, if the strict invariance test fails to provide acceptable data-model fit, then researchers may benefit from conducting the lower level tests and investigating the source of measurement non-invariance. Table 1 provides a summary of appropriate claims and comparisons at each level of measurement invariance.

For Practitioners

We encourage practitioners to use measurement invariance testing, when possible, in any endeavor to inform their practice where group comparisons with assessment instrument data of latent traits are utilized. Safeguarding against threats to the validity of the inferences drawn from group comparison studies is fundamental to the evaluation and success of inclusive pedagogies in the classroom. We acknowledge that sample size is often a limitation in many studies. Thus we advise practitioners to utilize similar processes of reflexivity to safeguard against threats to the validity of inferences against groups that are appropriate for their sample size, such as cognitive interviews (Willis, 1999). This practice will help to ensure that the investigations conducted across individual and institutional levels remain mindful of the tenets of CRT and move toward, rather than away from, equity. Additionally, we recommend the collaboration between practitioners and researchers in analyzing and interpreting quantitative data, particularly when comparing groups. These collaborations can be fruitful and inform a wider variety of settings in which our studies take place, providing the field of CER a broader and more complete view of the field as it advances toward greater diversity and social inclusion.

Lastly, we urge practitioners to review the research literature with a critical lens and hold research findings to a high standard when data is compared by group. Following the steps of measurement invariance testing can inform whether an instrument can be utilized to make meaningful comparisons with diverse groups. For a practical approach, if measurement

invariance testing is not feasible, we suggest a careful review of the literature for instruments which have been appropriately tested with diverse populations, to support appropriate data collection and analyses that lead to meaningful conclusions.

Conflicts of Interest

There are no conflicts to declare.

Acknowledgements

The authors wish to thank Gregory R. Hancock, Program Director of Measurement, Statistics and Evaluation and Director of the Center for Integrated Latent Variable Research (CILVR) at the University of Maryland for his thoughtful feedback about measurement invariance. Support was provided to G.A.R. by the National Science Foundation's Florida-Georgia Louis Stokes Alliance for Minority Participation Bridge to the Doctorate award 1612347. This material is also based upon work supported by the National Science Foundation under award 1849473 to J.E.L. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

AERA, APA, and NCME., (2014), *Standards for Educational and Psychological Testing,* American Psychological Association, Washington, DC.

Apple M. W., (2001), *Educating the 'Right' Way: Markets, Standards, God, and Inequality.* RoutledgeFalmer, New York, NY.

Arjoon J. A., Xu X., and Lewis J. E., (2013), Understanding the State of the Art for Measurement in Chemistry Education Research: Examining the Psychometric Evidence, *J. Chem. Educ.*, **90**, 536−545.

Beier M. E., Kim M. H., Saterbak A., Leautaud V., Bishnoi S., and Gilberto J. M., (2019), The effect of authentic project-based learning on attitudes and career aspirations in STEM, *J. Res. Sci. Teach.,* **56**(1), 3-23.

Bontempo D. E. and Hofer S. M., (2007), Assessing Factorial Invariance in Cross-Sectional and Longitudinal Studies., in Ong A. D. and van Dulmen M. H. M. (eds.), *Series in positive psychology. Oxford handbook of methods in positive psychology*. Oxford University Press, pp. 153–175.

Bornstein M. H., (1995), Form and function: Implications for studies of culture and human development, *Cult. Psychol.,* **1**(1), 123–137.

Bowen N. K., and Masa R. D., (2015), Conducting measurement invariance tests with ordinal data: A guide for social work researchers, *J. Soc. Social Work Res.,* **6**(2), 229-249.

Brandriet A. R., and Bretz S. L., (2014), The development of the redox concept inventory as a measure of students' symbolic and particular redox understandings and confidence, *J. Chem. Educ.,* **91**, 1132-1144.

Bretz S. L., (2014), Designing assessment tools to measure students' conceptual knowledge of chemistry. In *Tools of Chemistry Education Research,* Bunce D., Cole R., Eds., ACS Symposium Series.

Brown T. A., (2006), *Confirmatory Factor Analysis for Applied Research*, The Guilford Press, New York, NY.

Bunce D. M., Komperda R., Schroeder M. J., Dillner D. K., Lin S., Teichert M. A., and Hartman J. R., (2017), Differential use of study approaches by students of different achievement levels, *J. Chem. Educ.,* **94**(10), 1415–1424.

Byrne B. M., Shavelson R. J., and Muthén B., (1989), Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance, *Psychol. Bull.,* **105**(3), 456-466.

Candell G. L., and Drasgow F., (1988), An iterative procedure for linking metrics and assessing item bias in item response theory, *Appl. Psychol. Meas.,* **12**(3), 253-260.

Ceci S. J., Williams W. M., and Barnett S. M., (2009), Women's underrepresentation in science: Sociocultural and biological considerations, *Psychol. Bull.,* **135**(2), 218–261.

Chen F. F., (2007), Sensitivity of goodness of fit indexes to lack of measurement invariance, *Struct. Equ. Modeling*, **14**(3), 464-504.

Cheung G. W., and Rensvold R. B., (1999), Testing factorial invariance across groups: A reconceptualization and proposed new method, *J. Manage.,* **25**(1), 1-27.

Cheung G. W., and Rensvold R. B., (2002), Evaluating goodness-of-fit indexes for testing measurement invariance, *Struct. Equ. Modeling,* **9**(2), 233-255.

Cohen J., (1988), *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed.; Lawrence Erlbaum Associates: Hillsdale, NJ.

Counsell, A., Cribbie, R. A., and Flora, D. B., (2019), Evaluating equivalence testing methods for measurement invariance, *Multivar. Behav. Res.,* DOI:10.1080/00273171.2019.1633617

Covarrubias A., (2011), Quantitative intersectionality: A critical race analysis of the Chicana/o educational pipeline, *J. Latinos Educ.,* **10**(2), 86-105.

Covarrubias A., and Velez, V., (2013), Critical race quantitative intersectionality: An antiracists research paradigm that refuses to 'Let the numbers speak for themselves.' In *Handbook of Critical Race Theory in education,* (Eds.) Dixson A., Lynn, M. New York City, Routledge, pp. 270-285.

Crenshaw K., (1989), Demarginalizing the intersection of race and sex: A Black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics, *Univ. Chicago Leg. For.,* 139-168.

Crenshaw K., (1995), Critical Race Theory: The key writings that formed the movement, New York City: State University of New York Press.

Deng X., Doll W. J., Hendrickson A. R., and Scazzero J. A., (2005), A multi-group analysis of structural invariance: An illustration using the technology acceptance model, *Inform. Manage.,* **42***,* 745-759.

Delgado A., and Stefanic J., (2001), Critical Race Theory: An Introduction, New York City, NYU Press.

Dixson A., and Anderson C. R., (2018), Where are we? Critical Race Theory in education 20 years later, *Peabody J. Educ.,* **93**(1), 121-131.

Fernández L., (2002), Telling stories about school: Using critical race and Latino critical theories to document Latina/Latino education and resistance, *Qual. Inq.,* **8**(1), 45-65.

Ferrell B. and Barbera J., (2015), Analysis of students' self-efficacy, interest, and effort beliefs in general chemistry, *Chem. Educ. Res. Pract.,* **16**, 318-337.

Ferrell B., Phillips M. M., and Barbera J., (2016), Connecting achievement motivation to performance in general chemistry, *Chem. Educ. Res. Pract.,* **17**, 1054-1066.

Fink A., Cahill M. J., McDaniel M. A., Hoffman A., and Frey R. F., (2018), Improving general chemistry performance through a growth mindset intervention: selective effects on underrepresented minorities, *Chem. Educ. Res. Pract.,* **19***,* 783-806.

Finney S. J. and DiStefano C., (2013), Non-normal and categorical data in structural equation modeling., in Hancock G. R. and Mueller R. O. (eds.), *Structural equation modeling: a second course*. Charlotte, NC: Information Age Publishing, pp. 439–492.

Fischer R. and Karl J. A., (2019), A primer to (cross-cultural) multi-group invariance testing possibilities in R. *Front. Psychol.*, **10**, 1–18.

García N. M., López N., and Vélez V. N., (2018), QuantCrit: Rectifying quantitative methods through critical race theory, *Race Ethn. Educ.,* **21**(2), 149-157.

Garson D., (2012), *Testing statistical assumptions.* Statistical Associates Publishing, Asheboro, NC.

Gibbons R. E., and Raker J. R., (2018), Self-beliefs in organic chemistry: Evaluation of a reciprocal causation, cross-lagged model*, J. Res. Sci. Teach.*, **56**(5), 598-615.

Gibbons R. E., Xu X., Villafañe S. M., and Raker J. R., (2018), Testing a reciprocal causation model between anxiety, enjoyment and academic performance in postsecondary organic chemistry, *Educ. Psychol.*, **38** (6), 838-856.

Gillborn D., Warmington P., and Demack S., (2017), QuantCrit: Education, policy, 'big data' and principles for a critical race theory of statistics, *Race Ethn. Educ.,* **21** (2), 158-179.

Gregorich S. E., (2006), Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework, *Med Care,* **44** (11 Suppl 3), S78-S94.

Hancock G. R., (2001), Effect size, power, and sample size determination for structured means modeling and MIMIC approaches to between-groups hypothesis testing of means on a single latent construct, *Psychometrika,* **66**, 373–388.

Hancock G. R., and French B., (2013), *Power analysis in covariance structure modeling.* In G. R. Hancock and R. O. Mueller (Eds.), Structural equation modeling: A second course (2nd ed.), Information Age Publishing, Charlotte, NC, pp. 117-159.

Hancock G. R., Stapleton L. M., and Arnold-Berkovits I., (2009), The tenuousness of invariance tests within multisample covariance and mean structure models., in *Structural equation modeling in educational research: concepts and applications*., pp. 137–174.

Hensen C. and Barbera J. (2019), Assessing affective differences between a virtual general chemistry experiment and a similar hands-on experiment, *J. Chem. Educ.,* **96**, 2097−2108.

Hirschfeld G., and Von Brachel R., (2014), Multiple-Group confirmatory factor analysis in R – A tutorial in measurement invariance with continuous and ordinal, *Pract. Assess. Res. Eval.*, **19**(7), 1–11.

Hong L., and Page S. E., (2004), Groups of diverse problem solvers can outperform groups of high-ability problem solvers, *P.Natl. Acad. Sci.,* **101**(46), 16385-16389.

Hosbein K. N., and Barbera J., (2019), Evaluation of a novel measure of science and chemistry-specific identity, *Chem. Educ. Res. Pract.*, in review.

Hu L. T., and Bentler P. M., (1999), Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives, *Struct. Equ. Modeling,* **6**(1), 283−292.

Hurtado S., Newman C. B., Tran M. C., and Chang M. J., (2010), *Improving the Rate of Success for Underrepresented Racial Minorities in STEM Fields: Insights from a National Project.* In New Directions for Institutional Research, no. 148. Wiley Periodicals, Inc.

Ireland D. T., Freeman K .E., Winston-Proctor C. E., Delaine K. D., McDonald Lowe S., and Woodson K. M., (2018), (Un)hidden figures: A synthesis of research examining the intersectional experiences of Black women and girls in STEM, *Rev. Res. Educ.,* **42***,* 226-254.

Jiang B., Xu X., García A., and Lewis J. E., (2010), Comparing two tests of formal reasoning in a college chemistry context, *Chem. Educ. Res. Pract.,* **87**(12), 1430-1437.

Jöreskog K. G. (1971). Simultaneous factor analysis in several populations, *Psychometrika*, **36**, 409–426.

Kahveci A., (2015), Assessing high school students' attitudes toward chemistry with a shortened semantic differential, *Chem. Educ. Res. Pract.,* **16, 283–292.**

Kang Y., McNeish D. M., and Hancock G. R., (2016), The role of measurement quality on practical guidelines for assessing measurement and structural invariance. *Educ. Psychol. Meas.*, **76**(4), 533–561.

Keefer K. V., Holden R. R., and Parker J. D. A., (2013), Longitudinal assessment of trait emotional intelligence: Measurement invariance and construct continuity from late childhood to adolescence, *Psychol. Assess.,* **25**(4), 1255-1272.

Kendhammer L., Holme T., and Murphy K., (2013), Identifying differential performance in general chemistry: Differential item functioning analysis of ACS general chemistry trial tests, *J. Chem. Educ.,* **90***,* 846-853.

Kendhammer L. K., Murphy K., (2014), General statistical techniques for detecting differential item functioning based on gender subgroups: A comparison of the Mantel-Haenszel procedure, IRT, and logistic regression. In *Innovative Uses of Assessments for Teaching and Research ACS Symposium Series,* American Chemical Society: Washington, DC.

Komperda R., Hosbein K. N. and Barbera J., (2018), Evaluation of the influence of wording changes and course type on motivation instrument functioning in chemistry, *Chem. Educ. Res. Pract.,* **19**, 184-198.

Lieber R. L., (1990), Statistical significance and statistical power in hypothesis testing, *J. Orthop. Res.,* **8***,* 304-309.

Litzler E., Samuelson C. C., and Lorah J. A., (2014), Breaking it down: Engineering students STEM confidence at the intersection of race/ethnicity and gender, *Res. High. Educ.,* **55**, 810-832.

Liu Y., Ferrell B., Barbera J., and Lewis J. E., (2017), Development and evaluation of a chemistry-specific version of the academic motivation scale (AMS-Chem), *Chem. Educ. Res. Pract.,* **18***,* 191-213.

Loertscher J., (2010), Using assessment to improve learning in the biochemistry classroom, *Biochem. Mol. Biol. Educ.,* **38** (3), 188-189.

López N., Erwin C., Binder M., and Chavez M. J., (2018), Making the invisible visible: Advancing quantitative methods in higher education using Critical Race Theory and intersectionality, *Race Ethnic. Educ.,* **21**(2), 180-207.

McNeish D., An J., and Hancock G. R., (2018), The thorny relation between measurement quality and fit index cutoffs in latent variable models. *J. Pers. Assess.*, **100**(1), 43–52.

Mellenbergh G. J., (1989), Item bias and item response theory, *Int. J. Educ. Res.,* **13***,* 127-143.

Meredith W., (1993), Measurement equivalence, factor analysis, and factorial equivalence, *Psychometrika*, **58**, 525–543.

Messick S., (1995)., Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning, *Am. Psychol.,* **50**(9), 741-749.

Montes L. H., Ferreira R. A., and Rodriguez C., (2018), Explaining secondary school students' attitudes towards chemistry in Chile, *Chem. Educ. Res. Pract.,* **19***,* 533-542.

Mooring S. R., Mitchell C. E., and Burrows N. L., (2016), Evaluation of a flipped, large enrollment organic chemistry course on student attitude and achievement, *J. Chem. Educ.,* **93**, 1972-1883.

Mueller R. O., Hancock G. R., (2019), *Structural Equation Modeling.* In G. R. Hancock, L. M. Stapleton, and R. O. Mueller (Eds.), The reviewer's guide to quantitative methods in the social sciences (pp. 445-456). Routledge, New York, NY.

Muthén L. K. and Muthén B. O., (2010), *Mplus User's Guide*, 6th ed., Muthén and Muthén: Los Angeles, CA.

Narawathne I. N., (2019), Introducing diversity through an organic approach, *J. Chem. Educ.,* **96** (9), 2042-2049.

Narayanan A., (2012), A review of eight software packages for structural equation modeling. *Am. Stat.*, **66**(2), 129–138.

O'Shea S., Lysaght P., Roberts J., and Harwood V., (2016), Shifting the blame in higher education – social inclusion and deficit discourses, *High. Educ. Res. Dev.,* **35** (2), 322-336.

Puritty C., Strickland L. R., Alia E., Blonder B., Klein E., Kohl M. T., McGee E., Quintana M., Ridley R. E., Tellman B., and Gerber L. R., (2017), Without inclusion, diversity initiatives may not be enough, *Science,* **357** (6356), 1101-1102.

Putnick D. L., and Bornstein M. H., (2016), Measurement invariance conventions and reporting: The state of the art and future directions for psychological research, *Dev. Rev.,* **47***, 71–90.*
Rath K. A., Peterfreund A., Bayliss F., Runquist E., and Simonis U., (2012), Impact of supplemental instruction in entry-level chemistry courses at a midsized public university, *J. Chem. Educ.,* **89***,* 449-455.

Richards-Babb M., and Jackson J. K., (2011), Gendered responses to online homework use in general chemistry, *Chem. Educ. Res. Pract.,* **12***,* 409-419.

Roadrangka V., Yeany R. H., and Padilla M. J., (1983), Paper presented at the annual meeting of the National Association for Research in Science Teaching, Dallas, TX.

Rocabado G. A., Kilpatrick N. A., Mooring S. R., and Lewis J. E., (2019), Can we compare attitude scores among diverse populations? An exploration of measurement invariance testing to support valid comparisons between Black female students and their peers in an organic chemistry course, *J. Chem. Educ.,* **96**(11), 2371-2382.

Salta K., and Koulougliotis D., (2015), Assessing motivation to learn chemistry: Adaptation and validation of Science Motivation Questionnaire II with Greek secondary school students, *Chem. Educ. Res. Pract.,* **16,** 237-250.

Sass D., (2011), Testing measurement invariance and comparing latent factor means within a confirmatory factor analysis framework, *J. Psychoeduc. Assess.,* **29(4), 347–363.**

Seadler A., (2012), Obama introduces plan to increase U. S. STEM undergraduates, *Earth,* **57**(6), 27.

Shepard L. A., (1993), Evaluating Test Validity, *Am. Educ. Res. Assoc.,* **19,** 405-450.

Shortlidge E. E., Rain-Griffith L., Shelby C., Shusterman G. P., and Barbera J., (2019), Despite similar perceptions and attitudes, postbaccalaureate students outperform in introductory biology and chemistry courses, *CBE-Life Sci. Educ.,* **18**(3), 1-14.

Solórzano D. G., (1997), Images and words that wound: Critical Race Theory, racial stereotyping, and teacher education, *Teach. Educ. Quart.,* **24**(3), 5-19.

Solórzano D. G., (1998), Critical Race Theory , race and gender microaggressions, and the experiences of Chicana and Chicano scholars, *Int. J. Qual. Stud. Educ.,* **11**(1), 121-136.

Solórzano D. G., and Ornelas A., (2004), A critical race analysis of Latina/o and African American advanced placement enrollment in public high schools, *High School J.,* **87**(3), 15-26.

Stanich C. A., Pelch M. A., Theobald E. J., and Freeman S., (2018), A new approach to supplementary instruction narrows achievement and affect gaps for underrepresented minorities, first generation students, and women, *Chem. Res. Educ. Pract.,* **19***,* 846-866.

Steinmetz H., (2013), Analyzing Observed Composite Differences Across Groups. *Methodology*, **9**(1), 1–12.

Stevens J. P., (2007), *Intermediate Statistics: A Modern Approach*, 3rd Edition, Routledge Taylor and Francis Group, New York, NY.

Tobin K. G. and Capie W., (1981), The development and validation of a group test of logical thinking, *Educ. Psychol. Meas.*, **41**(2), 413–423.

Tsui L., (2007), Effective strategies to increase diversity in STEM fields: A review of the research literature, *J. Negro Educ.,* **76**(4), 555-581.

Vandenberg R. J., and Lance C. E., (2000), A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research, *Organ. Res. Methods,* **2**, 4–69.

Villafañe S. M., Bailey C. P., Loertscher J., Minderhout V., and Lewis J. E., (2011), Development and analysis of an instrument to assess student understanding of foundational concepts before biochemistry coursework, *Biochem. Mol. Biol. Educ.,* **39**(2), 102-109.

Villafañe S. M., García C. A., and Lewis J. E., (2014), Exploring diverse students' trends in chemistry self-efficacy throughout a semester of college-level preparatory chemistry, *Chem. Educ. Res. Pract.,* **15**(2), 114-127.

Wicherts J. M., Nolan C. V., and Heesen D. J., (2005), Stereotype threat and group differences in test performance: A question of measurement invariance, *J. Pers. Soc. Psychol.,* **89**(5), 686-716.

Widaman K. F., and Reise S. P., (1997), Exploring the measurement invariance of psychological instruments: Applications in the substance use domain, In: Bryant K. J., Windle M.E., West S.G., (Eds), *The Science of Prevention: Methodological Advances from Alcohol and Substance Abuse Research,* American Psychological Association, Washington, DC.

Willis G. B., (1999), Cognitive interviewing: A “how to” guide, *Meeting of the American Statistical Association,* Research Triangle Institute.

Wolf E. J., Harrington K. M., Clark S. L., and Miller M. W., (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety, *Educ. Psychol.,* **73**(6), 913-934.

Wren D., and Barbera J., (2013), Gathering evidence for validity during the design, development, and qualitative evaluation of the thermochemistry concept inventory, *J. Chem. Educ.,* **90***,* 1590-1601.

Xu X., Kim E. S., and Lewis J. E., (2016), Sex difference in spatial ability for college students and exploration of measurement invariance, *Learn. Individ. Differ.,* **45**, 176–184.

Xu X., Villafañe S. M., and Lewis J. E., (2013), College students’ attitudes toward chemistry, conceptual knowledge and achievement: structural equation model analysis, *Chem. Educ. Res. Pract.,* **14**(2), 188-200.

Yosso T., (2005), Whose culture has capital? *Race Ethnic.Educ.,* **8**(1), 69-91.

Electronic Supplementary Information

Addressing diversity and social inclusion through group comparisons: A primer on measurement invariance testing.

Guizella A. Rocabado,[1] Regis Komperda,[2‡] Jennifer E. Lewis[1,3] and Jack Barbera[4†]

The purpose of the electronic supplementary information (ESI) is to provide readers with the data and code necessary to reproduce the examples from the main body of the paper as well as to provide a template for conducting invariance testing on a simulated data set that can be modified for those interested in conducting invariance testing on their own data. The code in the ESI is primarily written for the R statistical computing language, though Mplus code is also included for conducting invariance testing. The code in the ESI is also available through GitHub (https://github.com/RegisBK/Invariance_CERP) as this provides an easier way to download and use the code rather than cutting and pasting from this document. All analyses were conducted with R version 3.6.1 (R Core Team, 2019) and Mplus version 8.2.

This document assumes a basic understanding of how to work with R and/or Mplus. Users less familiar with these programs are encouraged to consult any of the resources available describing the use of these programs (Hirschfeld and Von Brachel, 2014; Komperda, 2017; Muthén and Muthén, 2017; Rosseel, 2020). Unless otherwise noted, the code provided here is intended to be entered directly into the software and is written in a different font to distinguish it from explanatory text.

Table of Contents

[1] Department of Chemistry, University of South Florida.
[2] Department of Chemistry & Biochemistry; Center for Research in Mathematics and Science Education, San Diego State University.
[3] Center for the Improvement of Teaching and Research in Undergraduate STEM Education
[4] Department of Chemistry, Portland State University.
† Corresponding author for manuscript (jbarbera@pdx.edu)
‡ Corresponding author for ESI (rkomperda@sdsu.edu)
Electronic Supplementary Information (ESI) available: [ ] See DOI:

Simulation and Visualization of Data in R

Simulation of Identical Group Data

The data used for the examples in the main article are simulated data created in R to follow the structure of the fictional Perceived Relevance of Chemistry Questionnaire (PRCQ). The PRCQ is conceptualized as containing three fictitious subconstructs: Importance of Chemistry (IC), Connectedness of Chemistry (CC), and Applications of Chemistry (AC). Additionally, the fictitious PRCQ is designed to be a 12-item instrument, where there are four items designed to measure each of the three subconstructs. To simulate this data in R first requires the installation and loading of the package `simstandard` (Schneider, 2019) which requires other dependent packages such as `dplyr` (Wickham *et al.*, 2019) to be installed as well.

```
install.packages("simstandard")
library(simstandard)
```

Syntax from the `lavaan` factor analysis package (Rosseel, 2012) is used to specify a three-factor model with four items associated with each factor. For this model, named `PRCQ`, items 1–4 are associated with the IC factor, 5–8 with the CC factor, and 9–12 with the AC factor. All items are assigned to have the same strength of association with their respective factors, a standardized value of 0.8. This value was chosen as it is relatively strong but not perfect association. In addition, each factor was simulated as having a weak association with the other factors. IC and CC have an association of 0.3, IC and AC have an association of 0.2 and CC and AC have an association of 0.1.

```
PRCQ<-'
      IC =~ 0.8*I1 + 0.8*I2 + 0.8*I3 + 0.8*I4
      CC =~ 0.8*I5 + 0.8*I6 + 0.8*I7 + 0.8*I8
      AC =~ 0.8*I9 + 0.8*I10 + 0.8*I11 + 0.8*I12

      IC  =~  0.3*CC
      IC  =~  0.2*AC
      CC  =~  0.1*AC
'
```

Now, observed data that follow the relations described by the model can be simulated. The `set.seed()` function is used to ensure reproducibility across uses by simulating the same pseudorandom data each time the code is run. Following the example from the main text, data are simulated separately for 1000 fictional students in the STEM majors group and for 1000 students in the non-STEM majors group. A column named `group` is added to distinguish the data from each group and the two datasets are combined to form the new dataset named `combined`.

```
set.seed(1234)
STEM <- sim_standardized(PRCQ, n = 1000, observed = T, latent = F,
errors = F)
nonSTEM <- sim_standardized(PRCQ, n = 1000, observed = T, latent = F,
errors = F)
```

```
STEM$group<-"STEM"
nonSTEM$group<-"nonSTEM"

combined<-rbind(STEM, nonSTEM)
```

The data generated with `sim_standardized()` are standardized meaning they have an average value of 0 and standard deviation of 1 as well as a normal distribution. Descriptive statistics for the complete dataset and for each group within the dataset can be generated using the `describe()` and `describeBy()` functions in the `psych` package (Revelle, 2018) and are shown in Figure ESI1 and ESI2. Note that statistics are not generated for the `group` variable as it is a character, not a number.

```
library(psych)
describe(combined)
describeBy(combined, group="group")
```

```
> describe(combined)
       vars    n  mean   sd median trimmed  mad   min  max range  skew kurtosis   se
I1        1 2000 -0.01 0.98  -0.01   -0.01 0.97 -3.26 3.22  6.48  0.03    -0.02 0.02
I2        2 2000  0.00 1.00  -0.06    0.00 1.01 -3.44 3.44  6.88  0.02    -0.05 0.02
I3        3 2000 -0.03 0.99  -0.02   -0.03 0.98 -3.22 3.07  6.28  0.01    -0.05 0.02
I4        4 2000 -0.03 1.01   0.00   -0.03 1.02 -3.06 3.47  6.53  0.03    -0.12 0.02
I5        5 2000 -0.02 0.98  -0.01   -0.01 0.97 -3.13 3.39  6.52 -0.08    -0.06 0.02
I6        6 2000 -0.02 0.99  -0.01   -0.02 1.00 -3.53 3.36  6.89 -0.03     0.04 0.02
I7        7 2000  0.00 0.99   0.01    0.00 1.01 -3.03 4.22  7.24 -0.01     0.04 0.02
I8        8 2000  0.00 1.00   0.00    0.01 1.01 -3.63 3.03  6.66 -0.06    -0.03 0.02
I9        9 2000  0.01 0.97   0.03    0.01 0.97 -3.05 3.61  6.65  0.01     0.02 0.02
I10      10 2000  0.03 1.00   0.02    0.02 1.00 -2.95 3.40  6.35  0.11    -0.07 0.02
I11      11 2000  0.02 0.98   0.05    0.01 0.95 -2.88 3.48  6.36  0.04     0.00 0.02
I12      12 2000  0.01 0.98   0.01    0.00 0.98 -3.58 4.11  7.69  0.15     0.11 0.02
group*   13 2000   NaN   NA     NA     NaN   NA   Inf -Inf  -Inf    NA       NA   NA
```

Figure ESI1. Output from the `describe()` function using the dataset named `combined`.

```
> describeBy(combined, group="group")

 Descriptive statistics by group
group: nonSTEM
      vars    n  mean   sd median trimmed  mad   min  max range  skew kurtosis   se
I1       1 1000 -0.01 1.00  -0.04   -0.01 0.99 -3.26 3.15  6.42  0.06    -0.06 0.03
I2       2 1000 -0.02 1.01  -0.08   -0.02 1.00 -3.44 2.84  6.28 -0.03     0.02 0.03
I3       3 1000 -0.03 0.98  -0.01   -0.03 0.99 -2.90 3.07  5.97  0.01    -0.06 0.03
I4       4 1000 -0.05 1.02  -0.03   -0.05 1.06 -3.06 3.47  6.53  0.05    -0.17 0.03
I5       5 1000 -0.02 0.98  -0.01   -0.01 0.97 -3.13 3.39  6.52 -0.13     0.09 0.03
I6       6 1000 -0.05 1.01  -0.02   -0.04 1.03 -3.53 3.19  6.73 -0.12    -0.06 0.03
I7       7 1000 -0.04 1.01  -0.02   -0.03 1.04 -3.03 3.15  6.17 -0.05    -0.07 0.03
I8       8 1000 -0.01 1.02   0.00    0.00 1.00 -3.63 2.98  6.61 -0.08     0.12 0.03
I9       9 1000  0.04 0.97   0.06    0.04 0.97 -3.05 3.61  6.65  0.05     0.13 0.03
I10     10 1000  0.07 0.99   0.05    0.06 1.01 -2.74 3.08  5.82  0.12    -0.16 0.03
I11     11 1000  0.04 0.98   0.06    0.03 1.00 -2.66 3.35  6.00  0.10    -0.17 0.03
I12     12 1000  0.05 0.97   0.04    0.04 0.98 -3.58 4.11  7.69  0.15     0.35 0.03
group*  13 1000   NaN   NA     NA     NaN   NA   Inf -Inf  -Inf    NA       NA   NA
------------------------------------------------------------------------------------
group: STEM
      vars    n  mean   sd median trimmed  mad   min  max range  skew kurtosis   se
I1       1 1000  0.00 0.97   0.01    0.00 0.92 -3.02 3.22  6.23  0.00     0.01 0.03
I2       2 1000  0.01 0.99  -0.05    0.01 1.04 -2.74 3.44  6.18  0.07    -0.15 0.03
I3       3 1000 -0.03 1.00  -0.04   -0.03 0.99 -3.22 3.06  6.28  0.02    -0.06 0.03
I4       4 1000  0.00 1.00   0.00   -0.01 0.98 -3.05 3.21  6.26  0.02    -0.07 0.03
I5       5 1000 -0.01 0.98  -0.02   -0.01 0.99 -2.95 2.95  5.90 -0.03    -0.21 0.03
I6       6 1000  0.00 0.98  -0.01    0.00 0.96 -3.22 3.36  6.58  0.07     0.12 0.03
I7       7 1000  0.03 0.97   0.03    0.03 0.97 -2.91 4.22  7.12  0.05     0.13 0.03
I8       8 1000  0.01 0.98   0.02    0.02 1.02 -3.03 3.03  6.06 -0.03    -0.23 0.03
I9       9 1000 -0.02 0.97   0.00   -0.01 0.98 -2.79 2.89  5.68 -0.03    -0.11 0.03
I10     10 1000 -0.01 1.00  -0.03   -0.01 0.99 -2.95 3.40  6.35  0.10     0.00 0.03
I11     11 1000  0.00 0.97   0.04    0.00 0.90 -2.88 3.48  6.36 -0.02     0.15 0.03
I12     12 1000 -0.02 0.99  -0.03   -0.03 0.98 -2.67 3.14  5.81  0.15    -0.12 0.03
group*  13 1000   NaN   NA     NA     NaN   NA   Inf -Inf  -Inf    NA       NA   NA
```

Figure ESI2. Output by group from the `describeBy()` function using the dataset named `combined`.

Additionally, the data are complete with no missing cases. These data may not be representative of the type of data obtained in chemistry education research using a non-fictional assessment instrument. For the purposes of this example, as in the main body of the text, this dataset will continue to be used. Further procedures in the ESI will demonstrate converting the data from continuous into categorical, which may better match authentic data.

Simulation of Data with Unequal Factor Loadings and Unequal Item Means

The previous section described the simulation of data for two groups using the same model in each group. To illustrate the effect of invariance at different levels, modifications were made to the data. The data are simulated to highlight specific issues that could be encountered (i.e., noninvariant loadings, noninvariant intercepts) but are unlikely to be representative of authentic data which could have numerous issues simultaneously. The model below is used to simulate data with a lower association between AC and I10 for the non-STEM majors group (changed to 0.3 instead of 0.8), as used to generate Figure 4 in the manuscript. This data is combined with the original STEM majors data to create the `combined.invar.load` dataset.

```
PRCQ.invar.load<-'
      IC =~ 0.8*I1 + 0.8*I2 + 0.8*I3 + 0.8*I4
      CC =~ 0.8*I5 + 0.8*I6 + 0.8*I7 + 0.8*I8
      AC =~ 0.8*I9 + 0.3*I10 + 0.8*I11 + 0.8*I12

      IC  =~  0.3*CC
      IC  =~  0.2*AC
      CC  =~  0.1*AC
'

nonSTEM.invar.load <- sim_standardized(PRCQ.invar.load, n = 1000,
observed = T, latent = F, errors = F)

nonSTEM.invar.load$group<-"nonSTEM"

combined.invar.load<-rbind(STEM, nonSTEM.invar.load)
```

To create data with a higher mean for I3 in the STEM majors group, as used to generate Figures 4 and 5 in the manuscript, a new dataset is created from the original STEM majors data and constant of 2 is added to all values for I3 in this new data. The STEM majors data is combined with the original non-STEM majors data to create a `combined.invar.mean` dataset. The `describeBy()` function can be used to confirm differences between the groups as seen in the descriptive statistcs in Figure ESI3.

```
STEM.invar.mean<-STEM
STEM.invar.mean$I3<-STEM.invar.mean$I3+2

STEM.invar.mean$group<-"STEM"

combined.invar.mean<-rbind(STEM.invar.mean, nonSTEM)

describeBy(combined.invar.mean, group="group")
```

```
> describeBy(combined.invar.mean, group="group")

 Descriptive statistics by group
group: nonSTEM
       vars    n  mean   sd median trimmed  mad   min  max range  skew kurtosis   se
I1        1 1000 -0.01 1.00  -0.04   -0.01 0.99 -3.26 3.15  6.42  0.06    -0.06 0.03
I2        2 1000 -0.02 1.01  -0.08   -0.02 1.00 -3.44 2.84  6.28 -0.03     0.02 0.03
I3        3 1000 -0.03 0.98  -0.01   -0.03 0.99 -2.90 3.07  5.97  0.01    -0.06 0.03
I4        4 1000 -0.05 1.02  -0.03   -0.05 1.06 -3.06 3.47  6.53  0.05    -0.17 0.03
I5        5 1000 -0.02 0.98  -0.01   -0.01 0.97 -3.13 3.39  6.52 -0.13     0.09 0.03
I6        6 1000 -0.05 1.01  -0.02   -0.04 1.03 -3.53 3.19  6.73 -0.12    -0.06 0.03
I7        7 1000 -0.04 1.01  -0.02   -0.03 1.04 -3.03 3.15  6.17 -0.05    -0.07 0.03
I8        8 1000 -0.01 1.02   0.00    0.00 1.00 -3.63 2.98  6.61 -0.08     0.12 0.03
I9        9 1000  0.04 0.97   0.06    0.04 0.97 -3.05 3.61  6.65  0.05     0.13 0.03
I10      10 1000  0.07 0.99   0.05    0.06 1.01 -2.74 3.08  5.82  0.12    -0.16 0.03
I11      11 1000  0.04 0.98   0.06    0.03 1.00 -2.66 3.35  6.00  0.10    -0.17 0.03
I12      12 1000  0.05 0.97   0.04    0.04 0.98 -3.58 4.11  7.69  0.15     0.35 0.03
group*   13 1000   NaN   NA     NA     NaN   NA   Inf -Inf  -Inf    NA       NA   NA
------------------------------------------------------------------------------------
group: STEM
       vars    n  mean   sd median trimmed  mad   min  max range  skew kurtosis   se
I1        1 1000  0.00 0.97   0.01    0.00 0.92 -3.02 3.22  6.23  0.00     0.01 0.03
I2        2 1000  0.01 0.99  -0.05    0.01 1.04 -2.74 3.44  6.18  0.07    -0.15 0.03
I3        3 1000  1.97 1.00   1.96    1.97 0.99 -1.22 5.06  6.28  0.02    -0.06 0.03
I4        4 1000  0.00 1.00   0.00   -0.01 0.98 -3.05 3.21  6.26  0.02    -0.07 0.03
I5        5 1000 -0.01 0.98  -0.02   -0.01 0.99 -2.95 2.95  5.90 -0.03    -0.21 0.03
I6        6 1000  0.00 0.98  -0.01    0.00 0.96 -3.22 3.36  6.58  0.07     0.12 0.03
I7        7 1000  0.03 0.97   0.03    0.03 0.97 -2.91 4.22  7.12  0.05     0.13 0.03
I8        8 1000  0.01 0.98   0.02    0.02 1.02 -3.03 3.03  6.06 -0.03    -0.23 0.03
I9        9 1000 -0.02 0.97   0.00   -0.01 0.98 -2.79 2.89  5.68 -0.03    -0.11 0.03
I10      10 1000 -0.01 1.00  -0.03   -0.01 0.99 -2.95 3.40  6.35  0.10     0.00 0.03
I11      11 1000  0.00 0.97   0.04    0.00 0.90 -2.88 3.48  6.36 -0.02     0.15 0.03
I12      12 1000 -0.02 0.99  -0.03   -0.03 0.98 -2.67 3.14  5.81  0.15    -0.12 0.03
group*   13 1000   NaN   NA     NA     NaN   NA   Inf -Inf  -Inf    NA       NA   NA
```

Figure ESI3. Output by group from the `describeBy()` function using the dataset named `combined.invar.mean` showing different means for I3 across groups.

Visualization of Data

The R code in this section can be used to generate the data visualizations (correlations and distributions) shown in Figures 1–5 of the manuscript. Correlation plots can be made with the `corrplot` package (Wei and Simko, 2017). To use the `corrplot()` function, the numeric variables are selected from the `combined` dataset and a correlation matrix is generated with the `cor()` function. Additional function arguments are used to specify that colored boxes should be plotted (`method="color"`), the text should be in the diagonal of the matrix in black (`tl.pos="d", tl.col="black"`), only the lower diagonal of the correlation matrix should be visualized (`type="lower"`), and that grey grid lines should appear (`addgrid.col="grey"`). Specifying the size of the margins is done to make room for the plot title (`mar=c(0,0,1,0)`).

```
library(dplyr)
library(corrplot)

combined %>% select(I1:I12) %>%   cor() %>%
   corrplot(., method="color", tl.pos="d", tl.col="black",
   type="lower", addgrid.col="grey", mar=c(0,0,1,0))
```

Similar plots can be generated for subsets of the data by filtering the `combined` dataset using the group variable (`filter(group=="STEM")`).

```
combined %>% filter(group=="STEM") %>% select(I1:I12) %>% cor() %>%
   corrplot(., method="color", tl.pos="d", tl.col="black",
   type="lower", addgrid.col="grey", title="STEM Majors",
   mar=c(0,0,1,0))

combined %>% filter(group=="nonSTEM") %>% select(I1:I12) %>% cor() %>%
   corrplot(., method="color", tl.pos="d", tl.col="black",
   type="lower", addgrid.col="grey",title="Non-STEM Majors",
   mar=c(0,0,1,0))
```

Using the `combined.invar.load` dataset will produce Figure 3 images from the manuscript.

```
combined.invar.load %>% select(I1:I12) %>%   cor() %>%
   corrplot(., method="color", tl.pos="d", tl.col="black",
    type="lower", addgrid.col="grey",
    title="Combined Data Varied\n Strength of Association for I10",
   mar=c(0,0,1,0))

combined.invar.load %>% filter(group=="STEM") %>% select(I1:I12) %>%
   cor() %>% corrplot(., method="color", tl.pos="d", tl.col="black",
   type="lower", addgrid.col="grey", title="STEM Majors",
   mar=c(0,0,1,0))

combined.invar.load %>% filter(group=="nonSTEM")%>% select(I1:I12) %>%
   cor() %>% corrplot(., method="color", tl.pos="d", tl.col="black",
   type="lower", addgrid.col="grey",title="Non-STEM Majors",
   mar=c(0,0,1,0))
```

The Figure 4 images from the manucript are produced using the same method with the `combined.invar.mean` dataset.

```
combined.invar.mean %>% select(I1:I12) %>%   cor() %>%
   corrplot(., method="color", tl.pos="d", tl.col="black",
   type="lower", addgrid.col = "grey",
   title="Combined Data\n Varied Mean for I3",mar=c(0,0,1,0))

combined.invar.mean %>% filter(group=="STEM") %>% select(I1:I12) %>%
  cor() %>% corrplot(., method="color", tl.pos="d", tl.col="black",
  type="lower", addgrid.col = "grey", title="STEM Majors",
  mar=c(0,0,1,0))
```

```
combined.invar.mean %>% filter(group=="nonSTEM") %>% select(I1:I12)
   %>%  cor() %>% corrplot(., method="color", tl.pos="d",
   tl.col="black", type="lower", addgrid.col = "grey",
   title="Non-STEM Majors", mar=c(0,0,1,0))
```

In order to generate the boxplot Figure 5 of the manuscript the package `reshape2` (Wickham, 2007) is needed to restructure the dataset and the package `ggplot2` (Wickham, 2016) is used to create the plot. First, the STEM and non-STEM groups are given more descriptive names since those will appear in the figure legend. The groups are also ordered as with STEM Majors first since the default setting would put the groups in alphabetical order.

```
library(ggplot2)
library(reshape2)

combined.invar.mean$group<-ifelse(combined.invar.mean$group=="STEM",
                                  "STEM Majors", "Non-STEM Majors")
combined.invar.mean$group<-ordered(combined.invar.mean$group,
                                levels=c("STEM Majors", "Non-STEM Majors"))
```

Next, the `melt()` function is used to create a long-format dataset where each group, variable (Item), and value occupies a single column. This long format is necessary for plotting using the function `ggplot()` with `geom_boxplot()`. In this boxplot the x-axis is the group and the y-axis is the value for each variable (`x=group, y=value, fill=group`). Faceting by variable (`facet_grid(.~variable)`) plots each item separately, yet within a single plot. The remainder of the code provides graphical parameters.

```
melt.mean<-combined.invar.mean %>%
                         select(I1:I12, group) %>% melt(id="group")
melt.mean$group<-melt.mean$group %>% as.factor()

ggplot(melt.mean, aes(x=group, y=value, fill=group))+
  geom_boxplot() + facet_grid(.~variable) + theme_bw() +
  theme(axis.title.x=element_blank(), axis.text.x=element_blank(),
        axis.ticks.x=element_blank(), axis.title.y=element_blank(),
        legend.position="bottom") +
  scale_fill_discrete(name="Group")
```

Conducting Invariance Testing

This section provides an overview of how to conduct measurement invariance testing using two popular software platforms, R and Mplus. Results obtained from both pieces of software will be similar, so the selection of software depends on the preferences of the researcher. In addition to R and Mplus there are other tools available for conducting measurement invariance testing, including SAS, LISREL, EQS, or the AMOS add-in for SPSS. A helpful comparison of software for structural equation modeling with multiple groups can be found in Narayana (2012) and Byrne (2004) provides a guide to AMOS.

Before introducing the specific steps to take within R and Mplus, it is worthwhile to note the default settings of both software packages. Within R, the package `lavaan` is generally used for factor analyses and in this package the default way to provide scale to the factor is to fix the value of the first item loading to one. In Mplus, the factor is given scale by setting its variance to one. Both methods are acceptable ways of identifying the model and will give equivalent results. However, each of these methods has different implications in the context of measurement invariance testing with multiple groups.

The method of setting the factor variance to one (as in Mplus) in both groups is generally not recommended for multigroup measurement invariance testing as it implies that the latent variable has the same variance in both groups. This is described as homogeneity of variance for the latent variables. Though conceptually similar to the test for homogeneity of variance used in *t*-tests and ANOVAs, in a latent framework this is an untestable assumption (Hancock *et al.*, 2009, 168).

In the first method, used within `lavaan,` setting an item loading to one, the default is to use the first item on the scale. When the first item on the scale is set to be one for both groups the rest of the series of structural equations will be solved assuming this item has the same loading value in both groups. Yet, there is no way to know for certain if that assumption is true or if there are other scale items that would have been better to set equivalent. This seemingly inconsequential decision can have major implications for interpretation of results and researchers are advised to think carefully about which item may be best to set equal across groups based on either theoretical or observable grounds (Bontempo and Hofer, 2007; Hancock *et al.*, 2009).

Invariance Testing with R – Continuous Data

Within the R software, the package `lavaan`, previously used to generate the simulated data, can be used to test confirmatory factor (CFA) models as well as structural equation models (SEM). The function for performing CFA, `cfa()` contains built-in arguments to set various model parameters equal for invariance testing (Hirschfeld and Von Brachel, 2014), making invariance testing a relatively simple process. In this section, the steps for measurement invariance testing will follow those in the main article using the `combined.invar.load` dataset to generate the fit index data from Table 1 in the manuscript. The general process for invariance testing within R is that of building up from the least constrained model (i.e., configural invariance) to the most constrained model (i.e., conservative invariance). Identical steps can be followed for the other datasets and fit indices resulting from these tests are provided later sections.

Step 0: Establishing Baseline Model

Following the steps outline in the manuscript, the baseline model is tested for each group separately. The model is specified in the same manner as was used to generate the simulated data with the main difference being that values for the loadings and associations between factors are not assigned but will be estimated by the software from the data. This model is named `model.test` to distinguish it from the model used to simulate the data.

```
library(lavaan)

model.test<-'
      IC =~ I1 + I2 + I3 + I4
      CC =~ I5 + I6 + I7 + I8
      AC =~ I9 + I10 + I11 + I12
      '
```

The function `cfa()` is now used to examine how well the data fit the proposed model. The maximum likelihood (ML) estimator is used as the data are continuous and normally distributed and are therefore appropriate for the ML estimator. Additionally, this follows the steps in the main article and aligns with the estimator used to determine the suggested fit index cut off values (Hu and Bentler, 1999). In situations where the data are known to be nonnormally distributed the robust maximum likelihood estimator (MLR) is more appropriate and can be specified with the command `estimator="MLR"`. The results from ML and MLR are equivalent if the data are normal, and interested readers can confirm this for themselves since `lavaan` prints the output of both ML and MLR simultaneously when MLR is used. Later sections of this ESI will describe how to modify the code to accommodate categorical data. Finally, specify that the mean structure (intercepts) should be explicitly shown.

```
STEM.step0<-cfa(data=combined.invar.mean %>% filter(group=="STEM
                Majors"), model=model.test, estimator="ML",
                meanstructure=TRUE)
```

The `summary()` function provides a convenient way to view the fit statistics and model parameters from the model that was just fit to the STEM majors data.

```
summary(STEM.step0, standardized=TRUE, fit.measures=TRUE)
```

Though the output provided by `summary()` is extensive the key fit indices are indicated by boxes in Figure ESI4. Note that the fit indices match Table 1 in the manuscript and show essentially perfect fit: CFI > 0.95; SRMR < 0.08; RMSEA < 0.06 (Hu and Bentler, 1999).

```
> STEM.step0<-cfa(data = combined.invar.mean %>% filter(group=="STEM
Majors"), model = model.test, estimator="ML", meanstructure=TRUE)
> summary(STEM.step0, standardized=TRUE, fit.measures=TRUE)
lavaan 0.6-5 ended normally after 27 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of free parameters                         39

  Number of observations                          1000

Model Test User Model:

  Test statistic                                65.438
  Degrees of freedom                                51
  P-value (Chi-square)                           0.084

Model Test Baseline Model:

  Test statistic                              6052.309
  Degrees of freedom                                66
  P-value                                        0.000

User Model versus Baseline Model:

  Comparative Fit Index (CFI)                    0.998
  Tucker-Lewis Index (TLI)                       0.997

Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)             -13835.349
  Loglikelihood unrestricted model (H1)     -13802.630

  Akaike (AIC)                               27748.698
  Bayesian (BIC)                             27940.100
  Sample-size adjusted Bayesian (BIC)        27816.234

Root Mean Square Error of Approximation:

  RMSEA                                          0.017
  90 Percent confidence interval - lower         0.000
  90 Percent confidence interval - upper         0.028
  P-value RMSEA <= 0.05                          1.000

Standardized Root Mean Square Residual:

  SRMR                                           0.021
```

Figure ESI4. Summary output for testing baseline model (Step 0) with STEM majors data having modified I3 intercept highlighting chi square test statistic, degrees of freedom, *p*-value, CFI, RMSEA and SRMR.

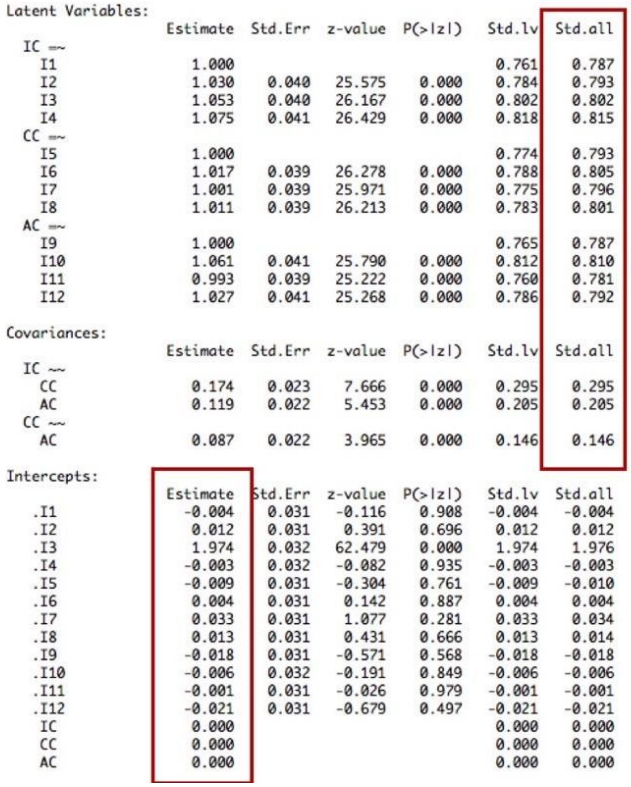
The same code can be executed using the non-STEM majors data and nearly identical fit is achieved (Figure ESI5).

```
nonSTEM.step0<-cfa(data=combined.invar.mean %>% filter(group=="Non-
                   STEM Majors"), model=model.test,
                   estimator="ML", meanstructure=TRUE)

summary(nonSTEM.step0, standardized=TRUE, fit.measures=TRUE)
```

```
> nonSTEM.step0<-cfa(data = combined.invar.mean %>% filter(group=="Non-
STEM Majors"), model = model.test, estimator="ML", meanstructure=TRUE)
> summary(nonSTEM.step0, standardized=TRUE, fit.measures=TRUE)
lavaan 0.6-5 ended normally after 30 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of free parameters                         39

  Number of observations                          1000

Model Test User Model:

  Test statistic                                51.931
  Degrees of freedom                               51
  P-value (Chi-square)                          0.437

Model Test Baseline Model:

  Test statistic                              6015.854
  Degrees of freedom                                66
  P-value                                        0.000

User Model versus Baseline Model:

  Comparative Fit Index (CFI)                    1.000
  Tucker-Lewis Index (TLI)                       1.000

Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)             -13981.961
  Loglikelihood unrestricted model (H1)     -13955.996

  Akaike (AIC)                               28041.922
  Bayesian (BIC)                             28233.325
  Sample-size adjusted Bayesian (BIC)        28109.459

Root Mean Square Error of Approximation:

  RMSEA                                          0.004
  90 Percent confidence interval - lower         0.000
  90 Percent confidence interval - upper         0.021
  P-value RMSEA <= 0.05                          1.000

Standardized Root Mean Square Residual:

  SRMR                                           0.016
```

Figure ESI5. R summary output for testing baseline model (Step 0) with unmodified non-STEM majors data highlighting chi square test statistic, degrees of freedom, *p*-value, CFI, RMSEA and SRMR.

Looking through the rest of the `summary()` output gives the values for the model parameters. The column `Std.all` is most typically reported when standardized model parameters are given. For both groups, these model parameters (Figures ESI6 & ESI7) match those used to simulate the data (loadings of 0.80 as well as associations between the three factors of approximately 0.3, 0.2, and 0.1). Examining the values of the intercept terms in both groups shows that in the STEM majors group (Figure ESI6) the intercept for I3 is larger than in the non-STEM majors group by a value of 2, as specified in the model used to simulate the data.



Latent Variables:

	Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
IC =~						
I1	1.000				0.761	0.787
I2	1.030	0.040	25.575	0.000	0.784	0.793
I3	1.053	0.040	26.167	0.000	0.802	0.802
I4	1.075	0.041	26.429	0.000	0.818	0.815
CC =~						
I5	1.000				0.774	0.793
I6	1.017	0.039	26.278	0.000	0.788	0.805
I7	1.001	0.039	25.971	0.000	0.775	0.796
I8	1.011	0.039	26.213	0.000	0.783	0.801
AC =~						
I9	1.000				0.765	0.787
I10	1.061	0.041	25.790	0.000	0.812	0.810
I11	0.993	0.039	25.222	0.000	0.760	0.781
I12	1.027	0.041	25.268	0.000	0.786	0.792

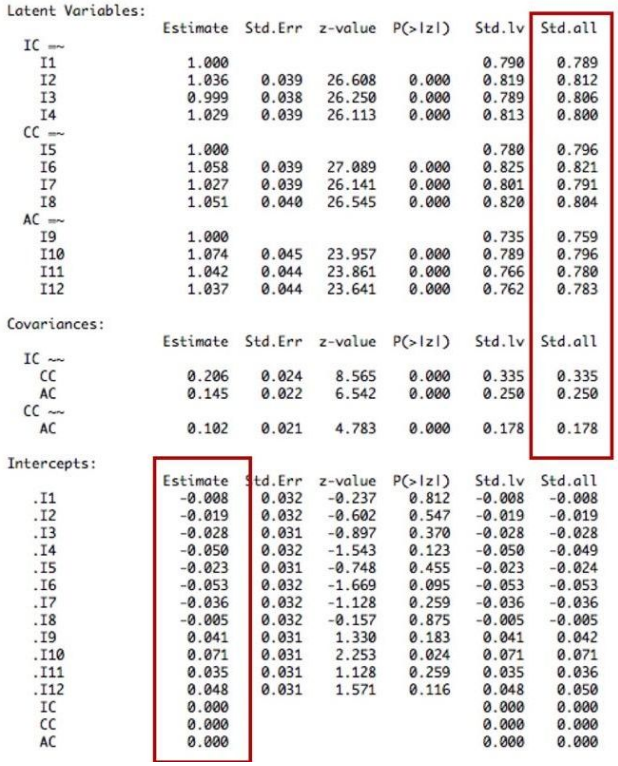
Covariances:

	Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
IC ~~						
CC	0.174	0.023	7.666	0.000	0.295	0.295
AC	0.119	0.022	5.453	0.000	0.205	0.205
CC ~~						
AC	0.087	0.022	3.965	0.000	0.146	0.146

Intercepts:

	Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
.I1	-0.004	0.031	-0.116	0.908	-0.004	-0.004
.I2	0.012	0.031	0.391	0.696	0.012	0.012
.I3	1.974	0.032	62.479	0.000	1.974	1.976
.I4	-0.003	0.032	-0.082	0.935	-0.003	-0.003
.I5	-0.009	0.031	-0.304	0.761	-0.009	-0.010
.I6	0.004	0.031	0.142	0.887	0.004	0.004
.I7	0.033	0.031	1.077	0.281	0.033	0.034
.I8	0.013	0.031	0.431	0.666	0.013	0.014
.I9	-0.018	0.031	-0.571	0.568	-0.018	-0.018
.I10	-0.006	0.032	-0.191	0.849	-0.006	-0.006
.I11	-0.001	0.031	-0.026	0.979	-0.001	-0.001
.I12	-0.021	0.031	-0.679	0.497	-0.021	-0.021
IC	0.000				0.000	0.000
CC	0.000				0.000	0.000
AC	0.000				0.000	0.000

Figure ESI6. R summary output for testing baseline model (Step 0) with unchanged STEM majors data highlighting standardized model parameters and intercepts.



Latent Variables:

	Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
IC =~						
I1	1.000				0.790	0.789
I2	1.036	0.039	26.608	0.000	0.819	0.812
I3	0.999	0.038	26.250	0.000	0.789	0.806
I4	1.029	0.039	26.113	0.000	0.813	0.800
CC =~						
I5	1.000				0.780	0.796
I6	1.058	0.039	27.089	0.000	0.825	0.821
I7	1.027	0.039	26.141	0.000	0.801	0.791
I8	1.051	0.040	26.545	0.000	0.820	0.804
AC =~						
I9	1.000				0.735	0.759
I10	1.074	0.045	23.957	0.000	0.789	0.796
I11	1.042	0.044	23.861	0.000	0.766	0.780
I12	1.037	0.044	23.641	0.000	0.762	0.783

Covariances:

	Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
IC ~~						
CC	0.206	0.024	8.565	0.000	0.335	0.335
AC	0.145	0.022	6.542	0.000	0.250	0.250
CC ~~						
AC	0.102	0.021	4.783	0.000	0.178	0.178

Intercepts:

	Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
.I1	-0.008	0.032	-0.237	0.812	-0.008	-0.008
.I2	-0.019	0.032	-0.602	0.547	-0.019	-0.019
.I3	-0.028	0.031	-0.897	0.370	-0.028	-0.028
.I4	-0.050	0.032	-1.543	0.123	-0.050	-0.049
.I5	-0.023	0.031	-0.748	0.455	-0.023	-0.024
.I6	-0.053	0.032	-1.669	0.095	-0.053	-0.053
.I7	-0.036	0.032	-1.128	0.259	-0.036	-0.036
.I8	-0.005	0.032	-0.157	0.875	-0.005	-0.005
.I9	0.041	0.031	1.330	0.183	0.041	0.042
.I10	0.071	0.031	2.253	0.024	0.071	0.071
.I11	0.035	0.031	1.128	0.259	0.035	0.036
.I12	0.048	0.031	1.571	0.116	0.048	0.050
IC	0.000				0.000	0.000
CC	0.000				0.000	0.000
AC	0.000				0.000	0.000

Figure ESI7. R summary output for testing baseline model (Step 0) with unchanged non-STEM majors data highlighting standardized model parameters and intercepts.

It is important to note that this difference in intercept for I3 between the groups (Figures ESI6 & ESI7) did not affect the overall fit of each group (Figures ESI4 & ESI5) because the parameters in each group were allowed to vary as needed to best fit the model. The purpose of testing these baseline models is to ensure that each group has a reasonable fit to the model before constraining any parameters to be equal across groups.

Step 1: Configural Invariance

The next step of invariance testing fits the model to both groups of data simultaneously. Within the `cfa()` function this is easily accomplished by specifying that groups are present and providing the name of the grouping variable (`group="group"`).

```
step1.comb.mean<-cfa(data=combined.invar.mean, model=model.test,
                     group="group", estimator="ML")

summary(step1.comb.mean, standardized=TRUE, fit.measures=TRUE)
```

Output from testing this model provides both an overall model chi square and the individual group chi square values obtained from Step 0 (Figure ESI8). The rest of the fit indices (CFI, RMSEA, and SRMR) are provided for the overall model. As show in Table 1 of the manuscript the fit indices for the configural model are essentially perfect. Further exploration of the model parameters shows that parameters for both groups have been estimated separately and match those in Step 0.

```
> step1.comb.mean<-cfa(data = combined.invar.mean, model = model.test,
group="group", estimator="ML")
> summary(step1.comb.mean, standardized=TRUE, fit.measures=TRUE)
lavaan 0.6-5 ended normally after 33 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of free parameters                         78

  Number of observations per group:
    STEM Majors                                   1000
    Non-STEM Majors                               1000

Model Test User Model:

  Test statistic                               117.369
  Degrees of freedom                               102
  P-value (Chi-square)                           0.142
  Test statistic for each group:
    STEM Majors                                 65.438
    Non-STEM Majors                             51.931

Model Test Baseline Model:

  Test statistic                             12068.162
  Degrees of freedom                               132
  P-value                                        0.000

User Model versus Baseline Model:

  Comparative Fit Index (CFI)                    0.999
  Tucker-Lewis Index (TLI)                       0.998

Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)             -27817.310
  Loglikelihood unrestricted model (H1)     -27758.626

  Akaike (AIC)                               55790.620
  Bayesian (BIC)                             56227.490
  Sample-size adjusted Bayesian (BIC)        55979.680

Root Mean Square Error of Approximation:

  RMSEA                                          0.012
  90 Percent confidence interval - lower         0.000
  90 Percent confidence interval - upper         0.021
  P-value RMSEA <= 0.05                          1.000

Standardized Root Mean Square Residual:

  SRMR                                           0.018
```

Figure ESI8. R summary output for configural invariance model (Step 1) with STEM majors data having modified I3 intercept highlighting chi square test statistic, degrees of freedom, *p*-value, CFI, RMSEA and SRMR.

Step 2: Metric Invariance (Weak)

To test for metric invariance (weak) the `group.equal` argument is used to specify that the loadings must be held constant across the two groups.

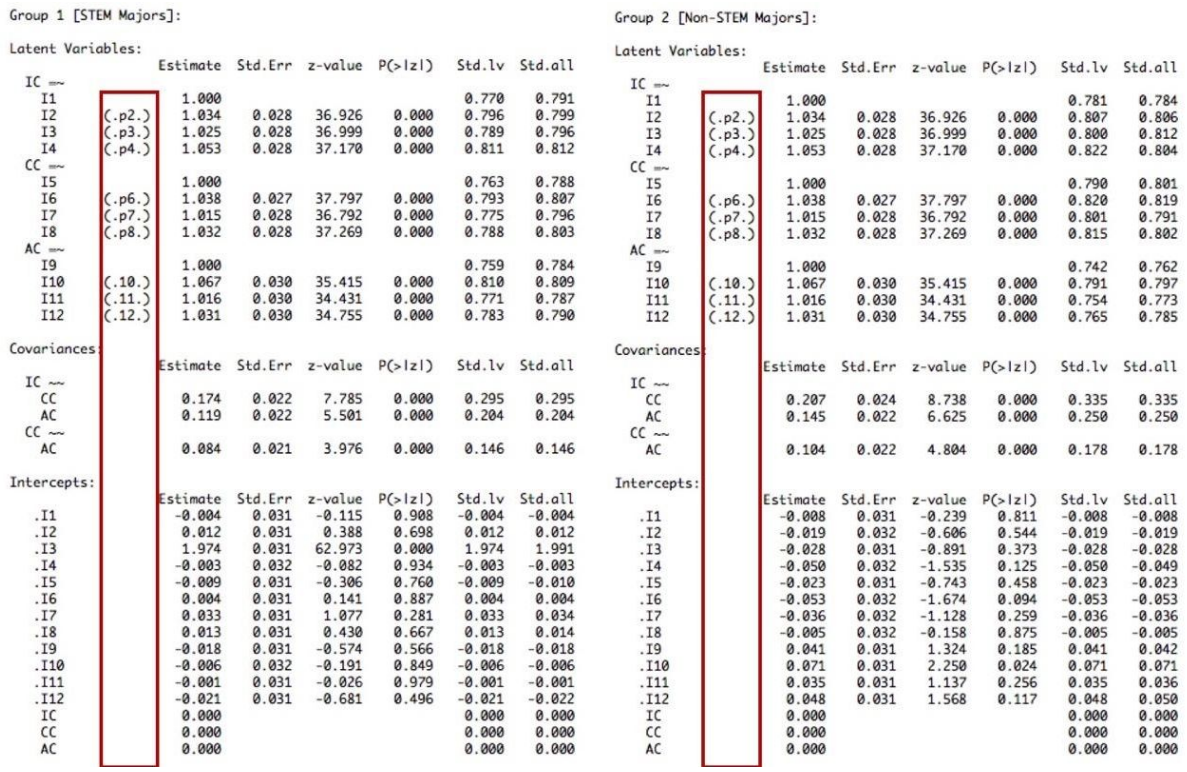
```
step2.comb.mean<-cfa(data=combined.invar.mean, model=model.test,
                     group="group", estimator="ML",
                     group.equal=c("loadings"))

summary(step2.comb.mean, standardized=TRUE, fit.measures=TRUE)
```

The fit indices for the metric invariance model (Figure ESI9) again match Table 1 in the manuscript and show essentially perfect fit. As described in the manuscript the change in fit index values can be calculated by hand but the *p*-value for the Δchi square must be computed.

```
> step2.comb.mean<-cfa(data = combined.invar.mean, model = model.test,
group="group", estimator="ML", group.equal=c("loadings"))
> summary(step2.comb.mean, standardized=TRUE, fit.measures=TRUE)
lavaan 0.6-5 ended normally after 30 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of free parameters                         78
  Number of equality constraints                     9
  Row rank of the constraints matrix                 9

  Number of observations per group:
    STEM Majors                                   1000
    Non-STEM Majors                               1000

Model Test User Model:

  Test statistic                               120.834
  Degrees of freedom                               111
  P-value (Chi-square)                           0.246
  Test statistic for each group:
    STEM Majors                                 67.162
    Non-STEM Majors                             53.672

Model Test Baseline Model:

  Test statistic                             12068.162
  Degrees of freedom                               132
  P-value                                        0.000

User Model versus Baseline Model:

  Comparative Fit Index (CFI)                    0.999
  Tucker-Lewis Index (TLI)                       0.999

Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)             -27819.043
  Loglikelihood unrestricted model (H1)     -27758.626

  Akaike (AIC)                               55776.085
  Bayesian (BIC)                             56162.547
  Sample-size adjusted Bayesian (BIC)        55943.331

Root Mean Square Error of Approximation:

  RMSEA                                          0.009
  90 Percent confidence interval - lower         0.000
  90 Percent confidence interval - upper         0.019
  P-value RMSEA <= 0.05                          1.000

Standardized Root Mean Square Residual:

  SRMR                                           0.019
```

Figure ESI9. R summary output for metric invariance model (Step 2) with STEM majors data having modified I3 intercept highlighting chi square test statistic, degrees of freedom, *p*-value, CFI, RMSEA and SRMR.

Examination of the model parameters is again done by groups (Figure ESI10) but shows that certain parameters have been constrained equal across the groups by assigning them a parameter name given in parenthesis (e.g., .p2.). Here the unstandardized loading values in the `Estimate` column are equal in both groups but the `Std.all` column values vary slightly. This is because the factors parameters (i.e., factor covariances) have not been constrained equal across groups and therefore affect the standardized loading values. Note that only the loadings have been assigned parameter names since these are the only parameters constrained equal across groups.



Group 1 [STEM Majors]:

Latent Variables:

		Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
IC =~							
I1		1.000				0.770	0.791
I2	(.p2.)	1.034	0.028	36.926	0.000	0.796	0.799
I3	(.p3.)	1.025	0.028	36.999	0.000	0.789	0.796
I4	(.p4.)	1.053	0.028	37.170	0.000	0.811	0.812
CC =~							
I5		1.000				0.763	0.788
I6	(.p6.)	1.038	0.027	37.797	0.000	0.793	0.807
I7	(.p7.)	1.015	0.028	36.792	0.000	0.775	0.796
I8	(.p8.)	1.032	0.028	37.269	0.000	0.788	0.803
AC =~							
I9		1.000				0.759	0.784
I10	(.10.)	1.067	0.030	35.415	0.000	0.810	0.809
I11	(.11.)	1.016	0.030	34.431	0.000	0.771	0.787
I12	(.12.)	1.031	0.030	34.755	0.000	0.783	0.790

Covariances:

	Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
IC ~~						
CC	0.174	0.022	7.785	0.000	0.295	0.295
AC	0.119	0.022	5.501	0.000	0.204	0.204
CC ~~						
AC	0.084	0.021	3.976	0.000	0.146	0.146

Intercepts:

	Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
.I1	-0.004	0.031	-0.115	0.908	-0.004	-0.004
.I2	0.012	0.031	0.388	0.698	0.012	0.012
.I3	1.974	0.031	62.973	0.000	1.974	1.991
.I4	-0.003	0.032	-0.082	0.934	-0.003	-0.003
.I5	-0.009	0.031	-0.306	0.760	-0.009	-0.010
.I6	0.004	0.031	0.141	0.887	0.004	0.004
.I7	0.033	0.031	1.077	0.281	0.033	0.034
.I8	0.013	0.031	0.430	0.667	0.013	0.014
.I9	-0.018	0.031	-0.574	0.566	-0.018	-0.018
.I10	-0.006	0.032	-0.191	0.849	-0.006	-0.006
.I11	-0.001	0.031	-0.026	0.979	-0.001	-0.001
.I12	-0.021	0.031	-0.681	0.496	-0.021	-0.022
IC	0.000				0.000	0.000
CC	0.000				0.000	0.000
AC	0.000				0.000	0.000

Group 2 [Non-STEM Majors]:

Latent Variables:

		Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
IC =~							
I1		1.000				0.781	0.784
I2	(.p2.)	1.034	0.028	36.926	0.000	0.807	0.806
I3	(.p3.)	1.025	0.028	36.999	0.000	0.800	0.812
I4	(.p4.)	1.053	0.028	37.170	0.000	0.822	0.804
CC =~							
I5		1.000				0.790	0.801
I6	(.p6.)	1.038	0.027	37.797	0.000	0.820	0.819
I7	(.p7.)	1.015	0.028	36.792	0.000	0.801	0.791
I8	(.p8.)	1.032	0.028	37.269	0.000	0.815	0.802
AC =~							
I9		1.000				0.742	0.762
I10	(.10.)	1.067	0.030	35.415	0.000	0.791	0.797
I11	(.11.)	1.016	0.030	34.431	0.000	0.754	0.773
I12	(.12.)	1.031	0.030	34.755	0.000	0.765	0.785

Covariances:

	Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
IC ~~						
CC	0.207	0.024	8.738	0.000	0.335	0.335
AC	0.145	0.022	6.625	0.000	0.250	0.250
CC ~~						
AC	0.104	0.022	4.804	0.000	0.178	0.178

Intercepts:

	Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
.I1	-0.008	0.031	-0.239	0.811	-0.008	-0.008
.I2	-0.019	0.032	-0.606	0.544	-0.019	-0.019
.I3	-0.028	0.031	-0.891	0.373	-0.028	-0.028
.I4	-0.050	0.032	-1.535	0.125	-0.050	-0.049
.I5	-0.023	0.031	-0.743	0.458	-0.023	-0.023
.I6	-0.053	0.032	-1.674	0.094	-0.053	-0.053
.I7	-0.036	0.032	-1.128	0.259	-0.036	-0.036
.I8	-0.005	0.032	-0.158	0.875	-0.005	-0.005
.I9	0.041	0.031	1.324	0.185	0.041	0.042
.I10	0.071	0.031	2.250	0.024	0.071	0.071
.I11	0.035	0.031	1.137	0.256	0.035	0.036
.I12	0.048	0.031	1.568	0.117	0.048	0.050
IC	0.000				0.000	0.000
CC	0.000				0.000	0.000
AC	0.000				0.000	0.000

Figure ESI10. R summary output for metric invariance model (Step 2) with STEM majors data having modified I3 intercept highlighting constraints on loading terms.

Step 3: Scalar Invariance (Strong)

Testing for scalar invariance only requires the addition of constraining the intercept terms to be equal, in addition to the loadings that were already constrained in Step 2.

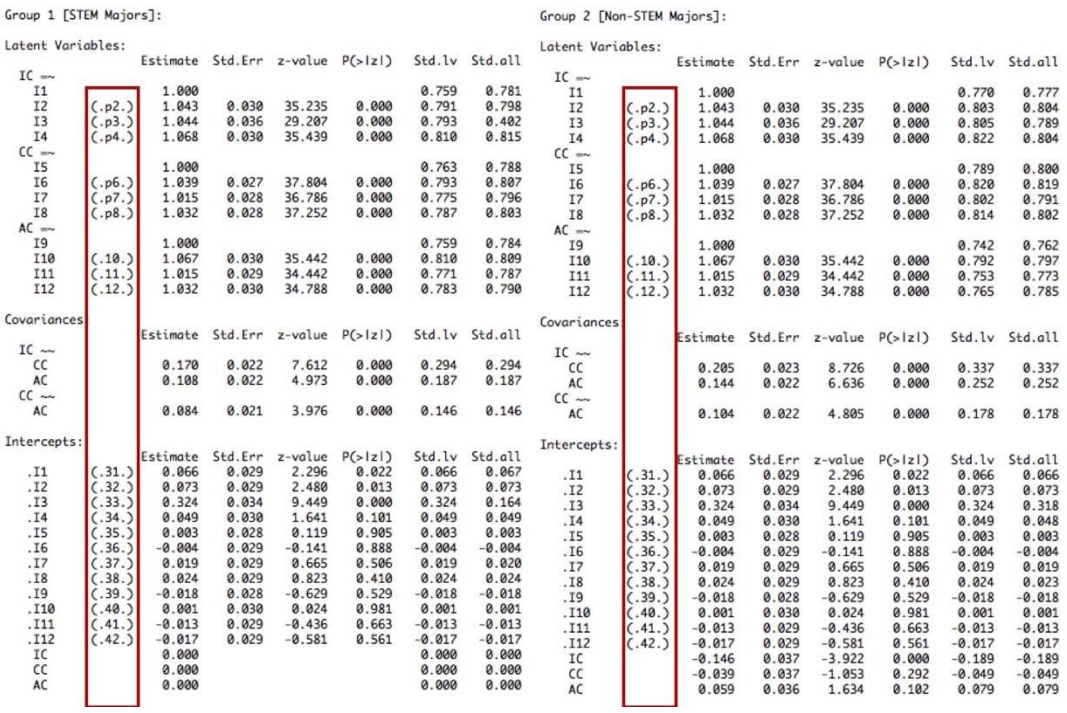
```
step3.comb.mean<-cfa(data=combined.invar.mean, model=model.test,
                     group="group", estimator="ML",
                     group.equal=c("loadings", "intercepts"))

summary(step3.comb.mean, standardized=TRUE, fit.measures=TRUE)
```

Again, matching the values found in Table 1 of the manuscript, the fit indices for the strict invariance model (Figure ESI11) indicate poor data-model fit, which is to be expected since the intercept terms were not simulated to be equal across groups. Notice that the chi square values for the individual groups give some indication that the problem is in the STEM Majors group, as

it has a much larger (worse) chi square value. Figure ESI12 shows that now the intercept terms are constrained to be equal across groups.

```
> step3.comb.mean<-cfa(data = combined.invar.mean, model = model.test,
group="group", estimator="ML", group.equal=c("loadings", "intercepts"))
> summary(step3.comb.mean, standardized=TRUE, fit.measures=TRUE)
lavaan 0.6-5 ended normally after 49 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of free parameters                         81
  Number of equality constraints                    21
  Row rank of the constraints matrix                21

  Number of observations per group:
    STEM Majors                                   1000
    Non-STEM Majors                               1000

Model Test User Model:

  Test statistic                              2267.834
  Degrees of freedom                           120
  P-value (Chi-square)                          0.000
  Test statistic for each group:
    STEM Majors                               2067.996
    Non-STEM Majors                            199.838

Model Test Baseline Model:

  Test statistic                             12068.162
  Degrees of freedom                           132
  P-value                                       0.000

User Model versus Baseline Model:

  Comparative Fit Index (CFI)                   0.820
  Tucker-Lewis Index (TLI)                      0.802

Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)            -28892.543
  Loglikelihood unrestricted model (H1)    -27758.626

  Akaike (AIC)                              57905.085
  Bayesian (BIC)                            58241.139
  Sample-size adjusted Bayesian (BIC)       58050.516

Root Mean Square Error of Approximation:

  RMSEA                                         0.134
  90 Percent confidence interval - lower        0.129
  90 Percent confidence interval - upper        0.139
  P-value RMSEA <= 0.05                         0.000

Standardized Root Mean Square Residual:

  SRMR                                          0.191
```

Figure ESI11. R summary output for metric invariance model (Step 3) with STEM majors data having modified I3 intercept highlighting chi square test statistic, degrees of freedom, *p*-value, CFI, RMSEA and SRMR.



Group 1 [STEM Majors]:

Latent Variables:

		Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
IC =~							
I1		1.000				0.759	0.781
I2	(.p2.)	1.043	0.030	35.235	0.000	0.791	0.798
I3	(.p3.)	1.044	0.036	29.207	0.000	0.793	0.402
I4	(.p4.)	1.068	0.030	35.439	0.000	0.810	0.815
CC =~							
I5		1.000				0.763	0.788
I6	(.p6.)	1.039	0.027	37.804	0.000	0.793	0.807
I7	(.p7.)	1.015	0.028	36.786	0.000	0.775	0.796
I8	(.p8.)	1.032	0.028	37.252	0.000	0.787	0.803
AC =~							
I9		1.000				0.759	0.784
I10	(.10.)	1.067	0.030	35.442	0.000	0.810	0.809
I11	(.11.)	1.015	0.029	34.442	0.000	0.771	0.787
I12	(.12.)	1.032	0.030	34.788	0.000	0.783	0.790

Covariances:

		Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
IC ~~							
CC		0.170	0.022	7.612	0.000	0.294	0.294
AC		0.108	0.022	4.973	0.000	0.187	0.187
CC ~~							
AC		0.084	0.021	3.976	0.000	0.146	0.146

Intercepts:

		Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
.I1	(.31.)	0.066	0.029	2.296	0.022	0.066	0.067
.I2	(.32.)	0.073	0.029	2.480	0.013	0.073	0.073
.I3	(.33.)	0.324	0.034	9.449	0.000	0.324	0.164
.I4	(.34.)	0.049	0.030	1.641	0.101	0.049	0.049
.I5	(.35.)	0.003	0.028	0.119	0.905	0.003	0.003
.I6	(.36.)	-0.004	0.029	-0.141	0.888	-0.004	-0.004
.I7	(.37.)	0.019	0.029	0.665	0.506	0.019	0.020
.I8	(.38.)	0.024	0.029	0.823	0.410	0.024	0.024
.I9	(.39.)	-0.018	0.028	-0.629	0.529	-0.018	-0.018
.I10	(.40.)	0.001	0.030	0.024	0.981	0.001	0.001
.I11	(.41.)	-0.013	0.029	-0.436	0.663	-0.013	-0.013
.I12	(.42.)	-0.017	0.029	-0.581	0.561	-0.017	-0.017
IC		0.000				0.000	0.000
CC		0.000				0.000	0.000
AC		0.000				0.000	0.000

Group 2 [Non-STEM Majors]:

Latent Variables:

		Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
IC =~							
I1		1.000				0.770	0.777
I2	(.p2.)	1.043	0.030	35.235	0.000	0.803	0.804
I3	(.p3.)	1.044	0.036	29.207	0.000	0.805	0.789
I4	(.p4.)	1.068	0.030	35.439	0.000	0.822	0.804
CC =~							
I5		1.000				0.789	0.800
I6	(.p6.)	1.039	0.027	37.804	0.000	0.820	0.819
I7	(.p7.)	1.015	0.028	36.786	0.000	0.802	0.791
I8	(.p8.)	1.032	0.028	37.252	0.000	0.814	0.802
AC =~							
I9		1.000				0.742	0.762
I10	(.10.)	1.067	0.030	35.442	0.000	0.792	0.797
I11	(.11.)	1.015	0.029	34.442	0.000	0.753	0.773
I12	(.12.)	1.032	0.030	34.788	0.000	0.765	0.785

Covariances:

		Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
IC ~~							
CC		0.205	0.023	8.726	0.000	0.337	0.337
AC		0.144	0.022	6.636	0.000	0.252	0.252
CC ~~							
AC		0.104	0.022	4.805	0.000	0.178	0.178

Intercepts:

		Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
.I1	(.31.)	0.066	0.029	2.296	0.022	0.066	0.066
.I2	(.32.)	0.073	0.029	2.480	0.013	0.073	0.073
.I3	(.33.)	0.324	0.034	9.449	0.000	0.324	0.318
.I4	(.34.)	0.049	0.030	1.641	0.101	0.049	0.048
.I5	(.35.)	0.003	0.028	0.119	0.905	0.003	0.003
.I6	(.36.)	-0.004	0.029	-0.141	0.888	-0.004	-0.004
.I7	(.37.)	0.019	0.029	0.665	0.506	0.019	0.019
.I8	(.38.)	0.024	0.029	0.823	0.410	0.024	0.023
.I9	(.39.)	-0.018	0.028	-0.629	0.529	-0.018	-0.018
.I10	(.40.)	0.001	0.030	0.024	0.981	0.001	0.001
.I11	(.41.)	-0.013	0.029	-0.436	0.663	-0.013	-0.013
.I12	(.42.)	-0.017	0.029	-0.581	0.561	-0.017	-0.017
IC		-0.146	0.037	-3.922	0.000	-0.189	-0.189
CC		-0.039	0.037	-1.053	0.292	-0.049	-0.049
AC		0.059	0.036	1.634	0.102	0.079	0.079

Figure ESI12. R summary output for scalar invariance model (Step 3) with STEM majors data having modified I3 intercept highlighting constraints on loading and intercept terms.

Step 4: Conservative Invariance (Strict)

Given the poor fit of the scalar invariance model, and out of range delta fit index values, it is not appropriate to go on to consider the strict invariance model. However, interested readers can test this model by adding "`residuals`" to the `group.equal` argument (residuals is another name for the error variance terms).

```
Step4.comb.mean<-cfa(data=combined.invar.mean, model=model.test,
                     group="group", estimator="ML",
                     group.equal=c("loadings", "intercepts", "residuals"))

summary(step4.comb.mean, standardized=TRUE, fit.measures=TRUE)
```

Exporting Data from R to Mplus

Data within R can be exported in a variety of familiar formats including txt, csv, and xlsx. Most conveniently for those working in Mplus there is also a package, `MplusAutomation` (Hallquist and Wiley, 2018), that allows for direct export of data in the correct Mplus format, dat. The correct format for Mplus requires data to not have any header information, such as column names. The `MplusAutomation` package also generates appropriate code to communicate the structure of the file to Mplus. The R code below shows how to export the simulated PRCQ data to Mplus and request the input file, which provides the code to use within Mplus to import the dat file in the correct format to be read by Mplus. Note that the group variable had been stored as a categorical factor within R and must be changed to a numeric variable for export. In this case the first group (STEM majors) will become 1 and the second group will become 2. This can be confirmed with the `describeBy()` function.

```
library(MplusAutomation)

combined.invar.mean$group<-combined.invar.mean$group %>%
                            as.numeric()
describeBy(combined.invar.mean, group="group")

prepareMplusData(combined.invar.mean,
                 filename="InvarianceMean.dat", inpfile = TRUE,
                 keepCols=c("I1", "I2", "I3", "I4","I5", "I6",
                 "I7", "I8", "I9", "I10", "I11", "I12", "group"))
```

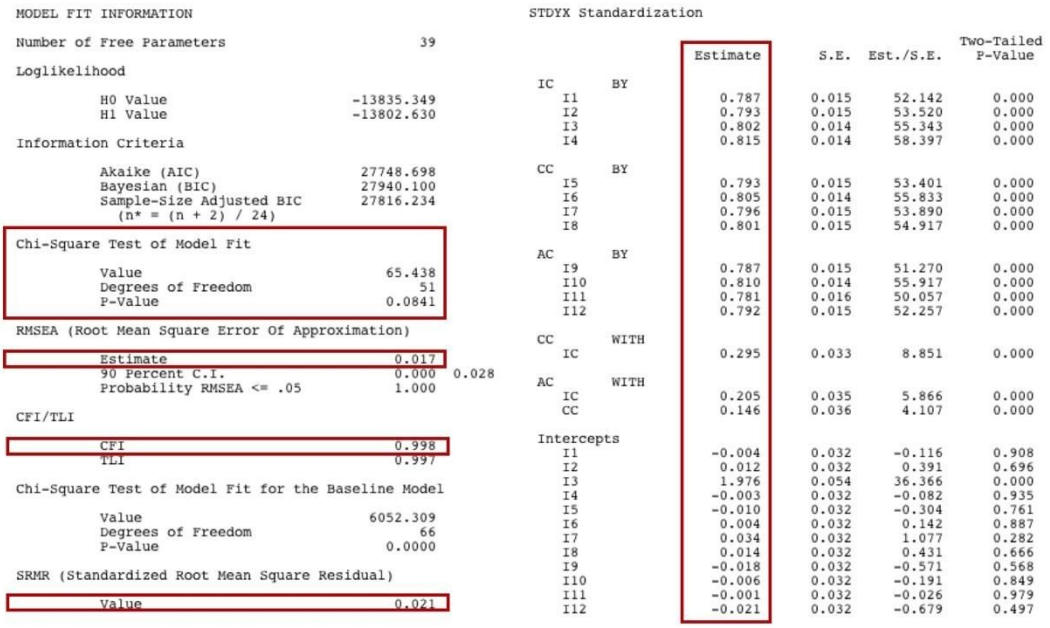
As a result of these commands R will create two new files, `InvarianceMean.dat` and `InvarianceMean.inp` in the working directory of your R session. If you are unsure of where your working directory resides, use the command `getwd()`.

Invariance Testing with Mplus – Continuous Data

Invariance testing in Mplus begins by opening the `inp` file generated previously or creating a new `inp` file for your own data. At the top of the `inp` file will be a title for the model being tested, the name of the data file, and the names of the variables in the data file. As before, the first step should be to test the model for each group individual. This is accomplished with the command `USEOBSERVATIONS`. Then the model to be tested is specified, this step is similar to lavaan but uses the term `BY` instead of =~ to denote relations between items and factors.

```
TITLE: STEM Majors Group Step 0
DATA: FILE = "InvarianceMean.dat";
VARIABLE:
NAMES = I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12 group;
USEVARIABLES ARE I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12;
USEOBSERVATIONS are group==1;

MODEL:
IC BY I1 I2 I3 I4;
CC BY I5 I6 I7 I8;
AC BY I9 I10 I11 I12;

OUTPUT:
STANDARDIZED;
```

The output for this model provides the same fit indices and standardized model parameters (Figure ESI13) as produced in R (Figures ESI4 & ESI 6) and shown in Table 1 of the manuscript.



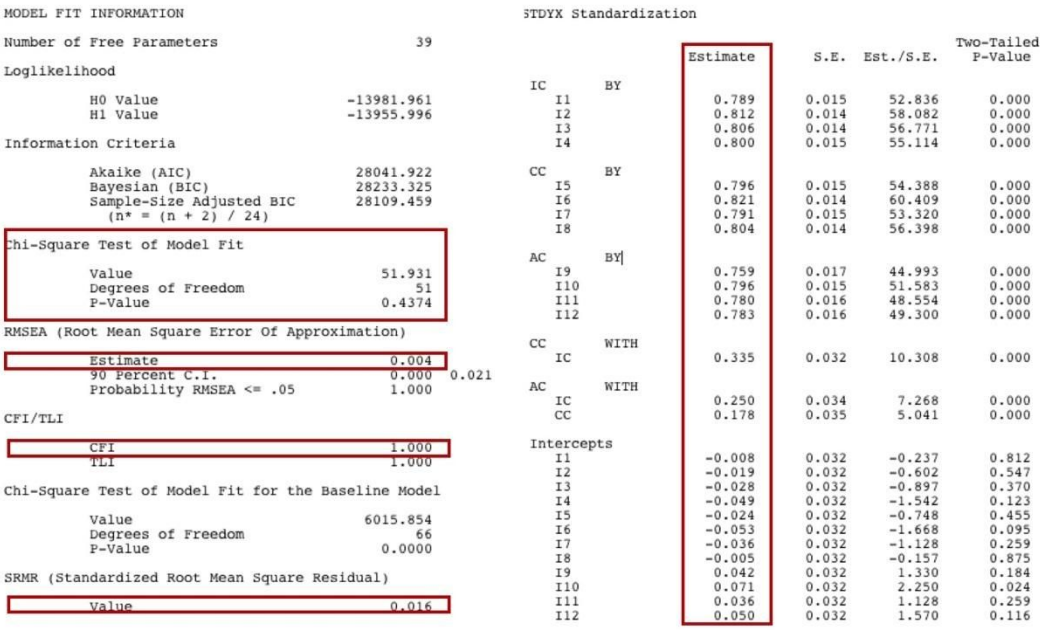
```
MODEL FIT INFORMATION

Number of Free Parameters                       39

Loglikelihood

          H0 Value                       -13835.349
          H1 Value                       -13802.630

Information Criteria

          Akaike (AIC)                    27748.698
          Bayesian (BIC)                  27940.100
          Sample-Size Adjusted BIC        27816.234
            (n* = (n + 2) / 24)

Chi-Square Test of Model Fit

          Value                              65.438
          Degrees of Freedom                    51
          P-Value                           0.0841

RMSEA (Root Mean Square Error Of Approximation)

          Estimate                           0.017
          90 Percent C.I.                    0.000  0.028
          Probability RMSEA <= .05           1.000

CFI/TLI

          CFI                                0.998
          TLI                                0.997

Chi-Square Test of Model Fit for the Baseline Model

          Value                           6052.309
          Degrees of Freedom                    66
          P-Value                           0.0000

SRMR (Standardized Root Mean Square Residual)

          Value                              0.021
```

```
STDYX Standardization

                                                        Two-Tailed
                    Estimate       S.E.  Est./S.E.    P-Value
IC       BY
   I1                  0.787      0.015     52.142      0.000
   I2                  0.793      0.015     53.520      0.000
   I3                  0.802      0.014     55.343      0.000
   I4                  0.815      0.014     58.397      0.000
CC       BY
   I5                  0.793      0.015     53.401      0.000
   I6                  0.805      0.014     55.833      0.000
   I7                  0.796      0.015     53.890      0.000
   I8                  0.801      0.015     54.917      0.000
AC       BY
   I9                  0.787      0.015     51.270      0.000
   I10                 0.810      0.014     55.917      0.000
   I11                 0.781      0.016     50.057      0.000
   I12                 0.792      0.015     52.257      0.000
CC       WITH
   IC                  0.295      0.033      8.851      0.000
AC       WITH
   IC                  0.205      0.035      5.866      0.000
   CC                  0.146      0.036      4.107      0.000
Intercepts
   I1                 -0.004      0.032     -0.116      0.908
   I2                  0.012      0.032      0.391      0.696
   I3                  1.976      0.054     36.366      0.000
   I4                 -0.003      0.032     -0.082      0.935
   I5                 -0.010      0.032     -0.304      0.761
   I6                  0.004      0.032      0.142      0.887
   I7                  0.034      0.032      1.077      0.282
   I8                  0.014      0.032      0.431      0.666
   I9                 -0.018      0.032     -0.571      0.568
   I10                -0.006      0.032     -0.191      0.849
   I11                -0.001      0.032     -0.026      0.979
   I12                -0.021      0.032     -0.679      0.497
```

Figure ESI13. Mplus summary output baseline model (Step 0) with STEM majors data having modified I3 intercept highlighting chi square test statistic, degrees of freedom, *p*-value, CFI, RMSEA, SRMR, and standardized model parameters.

Similar code can be used for the non-STEM majors group and again the results (Figure ESI14) will agree with the R output (Figures ESI15 & ESI17 as well as Table 1 of the manuscript.

```
TITLE: Non-STEM Majors Group Step 0
DATA: FILE = "InvarianceMean.dat";
VARIABLE:
NAMES = I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12 group;
USEVARIABLES ARE I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12;
USEOBSERVATIONS are group==2;

MODEL:
IC BY I1 I2 I3 I4;
CC BY I5 I6 I7 I8;
AC BY I9 I10 I11 I12;

OUTPUT:
STANDARDIZED;
```



```
MODEL FIT INFORMATION

Number of Free Parameters                       39

Loglikelihood

          H0 Value                       -13981.961
          H1 Value                       -13955.996

Information Criteria

          Akaike (AIC)                    28041.922
          Bayesian (BIC)                  28233.325
          Sample-Size Adjusted BIC        28109.459
            (n* = (n + 2) / 24)

Chi-Square Test of Model Fit

          Value                              51.931
          Degrees of Freedom                     51
          P-Value                            0.4374

RMSEA (Root Mean Square Error Of Approximation)

          Estimate                           0.004
          90 Percent C.I.                    0.000  0.021
          Probability RMSEA <= .05           1.000

CFI/TLI

          CFI                                1.000
          TLI                                1.000

Chi-Square Test of Model Fit for the Baseline Model

          Value                           6015.854
          Degrees of Freedom                   66
          P-Value                            0.0000

SRMR (Standardized Root Mean Square Residual)

          Value                              0.016
```

```
STDYX Standardization

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

IC         BY
    I1                 0.789      0.015     52.836      0.000
    I2                 0.812      0.014     58.082      0.000
    I3                 0.806      0.014     56.771      0.000
    I4                 0.800      0.015     55.114      0.000

CC         BY
    I5                 0.796      0.015     54.388      0.000
    I6                 0.821      0.014     60.409      0.000
    I7                 0.791      0.015     53.320      0.000
    I8                 0.804      0.014     56.398      0.000

AC         BY
    I9                 0.759      0.017     44.993      0.000
    I10                0.796      0.015     51.583      0.000
    I11                0.780      0.016     48.554      0.000
    I12                0.783      0.016     49.300      0.000

CC         WITH
    IC                 0.335      0.032     10.308      0.000

AC         WITH
    IC                 0.250      0.034      7.268      0.000
    CC                 0.178      0.035      5.041      0.000

Intercepts
    I1                -0.008      0.032     -0.237      0.812
    I2                -0.019      0.032     -0.602      0.547
    I3                -0.028      0.032     -0.897      0.370
    I4                -0.049      0.032     -1.542      0.123
    I5                -0.024      0.032     -0.748      0.455
    I6                -0.053      0.032     -1.668      0.095
    I7                -0.036      0.032     -1.128      0.259
    I8                -0.005      0.032     -0.157      0.875
    I9                 0.042      0.032      1.330      0.184
    I10                0.071      0.032      2.250      0.024
    I11                0.036      0.032      1.128      0.259
    I12                0.050      0.032      1.570      0.116
```

Figure ESI14. Mplus summary output baseline model (Step 0) with Non-STEM majors data having modified I3 intercept highlighting chi square test statistic, degrees of freedom, *p*-value, CFI, RMSEA, SRMR, and standardized model parameters.

Step 1: Configural Invariance

To test configural invariance within Mplus, the model is specified separately for each group. The `!` notation is used to insert comments within the Mplus model code. To provide results aligned with the R output the `@1` notation is used to identify the model by standardizing the loading for the first item on each factor. This is the default setting for the R `cfa()` function, but models in both programs can also be run by standardizing the factors instead of the loadings as a method of identifying the model.

Next the factor intercept is set to zero using brackets and `@0` notation. By default, Mplus assumes that item intercepts should be equal across groups, these can be freely estimated using the bracket notation. Item error variances are coded without the use of brackets. Specifying the same model for the second group will tell Mplus to estimate parameters for both models separately.

```
TITLE: Combined Dataset with Mean Differences Step 1 (Configural)
DATA: FILE = "InvarianceMean.dat";
VARIABLE:
NAMES = I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12 group;
USEVARIABLES ARE I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12;
GROUPING = group (1 = STEM 2 = NonSTEM);

MODEL:
! Model with standardized loading of first item on each factor
  IC BY I1@1 I2 I3 I4;
  CC BY I5@1 I6 I7 I8;
  AC BY I9@1 I10 I11 I12;

! Setting factor intercepts to zero
  [IC@0];
  [CC@0];
  [AC@0];

! Allowing item intercepts to be freely estimated
  [I1-I12];

! Allowing item error variances to be freely estimated
  I1-I12;

! Specifying the same model for the second group will cause
! all parameters to be freely estimated for the second group
MODEL NonSTEM:
  IC BY I1@1 I2 I3 I4;
  CC BY I5@1 I6 I7 I8;
  AC BY I9@1 I10 I11 I12;

  [IC@0];
  [CC@0];
  [AC@0];

  [I1-I12];

  I1-I12;
OUTPUT:
STANDARDIZED;
```

The output from this model (Figure ESI15) matches the fit indices in Table 1 of the manuscript for the configural model and both the unstandardized and standardized model parameters for the STEM majors group (Figure ESI16) and non-STEM majors group match those found using R (Figures ESI6 & ESI7).

```
MODEL FIT INFORMATION

Number of Free Parameters                       78

Loglikelihood

          H0 Value                       -27817.310
          H1 Value                       -27758.626

Information Criteria

          Akaike (AIC)                    55790.620
          Bayesian (BIC)                  56227.490
          Sample-Size Adjusted BIC        55979.680
            (n* = (n + 2) / 24)

Chi-Square Test of Model Fit

          Value                             117.369
          Degrees of Freedom                    102
          P-Value                            0.1418

Chi-Square Contribution From Each Group

          STEM                               65.438
          NONSTEM                            51.931

RMSEA (Root Mean Square Error Of Approximation)

          Estimate                           0.012
          90 Percent C.I.                    0.000  0.021
          Probability RMSEA <= .05           1.000

CFI/TLI

          CFI                                0.999
          TLI                                0.998

Chi-Square Test of Model Fit for the Baseline Model

          Value                          12068.162
          Degrees of Freedom                   132
          P-Value                            0.0000

SRMR (Standardized Root Mean Square Residual)

          Value                              0.019
```

Figure ESI15. Mplus summary output for configural invariance (Step 1) with STEM majors data having modified I3 intercept highlighting fit information.

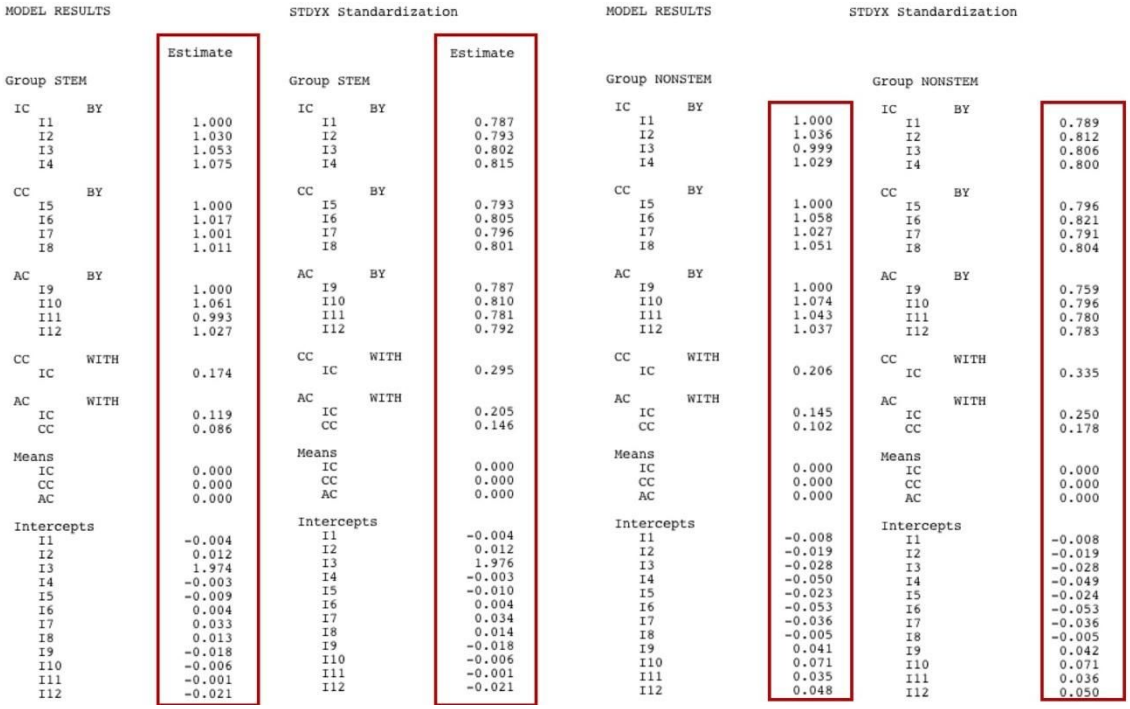


Figure ESI16. Mplus output for configural invariance (Step 1) with STEM majors data having modified I3 intercept highlighting unstandardized and standardized model parameters for both groups.

Step 2: Metric Invariance (Weak)

Metric invariance is tested by assigning the same parameter names to the loading terms in each group. In this example the names `L1-L12` are assigned to each of the loading parameters. Repeating this assignment in the second group will cause Mplus to set the unstandardized value of the parameters equal.

```
TITLE: Combined Dataset with Mean Differences Step 2 (Weak)
DATA: FILE = "InvarianceMean.dat";
VARIABLE:
NAMES = I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12 group;
USEVARIABLES ARE I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12;
GROUPING = group (1 = STEM 2 = NonSTEM);

MODEL:
! Model with standardized loading of first item on each factor
! Assigning a parameter name to each loading value (L1-L12)
  IC BY I1@1 I2 I3 I4 (L1-L4);
  CC BY I5@1 I6 I7 I8 (L5-L8);
  AC BY I9@1 I10 I11 I12 (L9-L12);

! Setting factor intercepts to zero
  [IC@0];
  [CC@0];
  [AC@0];

! Allowing item intercepts to be freely estimated
  [I1-I12];

! Allowing item error variances to be freely estimated
  I1-I12;

! Specifying the same model for the second group will force
! loadings to be equivalent across groups while other
! parameters are freely estimated
MODEL NonSTEM:
  IC BY I1@1 I2 I3 I4 (L1-L4);
  CC BY I5@1 I6 I7 I8 (L5-L8);
  AC BY I9@1 I10 I11 I12 (L9-L12);

  [IC@0];
  [CC@0];
  [AC@0];

  [I1-I12];

  I1-I12;

OUTPUT:
STANDARDIZED;
```

The output from this model (Figure ESI17) matches the fit indices in Table 1 of the manuscript for the weak invariance model and now the unstandardized parameters are equal across groups (Figure ESI18) while the intercepts are allowed to differ. As before, the standardized parameters differ slightly, but are aligned with the R output (Figure ESI10).

```
MODEL FIT INFORMATION

Number of Free Parameters                       69

Loglikelihood

          H0 Value                       -27819.043
          H1 Value                       -27758.626

Information Criteria

          Akaike (AIC)                    55776.085
          Bayesian (BIC)                  56162.547
          Sample-Size Adjusted BIC        55943.331
            (n* = (n + 2) / 24)

Chi-Square Test of Model Fit

          Value                             120.834
          Degrees of Freedom                    111
          P-Value                            0.2464

Chi-Square Contribution From Each Group

          STEM                               67.159
          NONSTEM                            53.674

RMSEA (Root Mean Square Error Of Approximation)

          Estimate                           0.009
          90 Percent C.I.                    0.000  0.019
          Probability RMSEA <= .05           1.000

CFI/TLI

          CFI                                0.999
          TLI                                0.999

Chi-Square Test of Model Fit for the Baseline Model

          Value                          12068.162
          Degrees of Freedom                    132
          P-Value                            0.0000

SRMR (Standardized Root Mean Square Residual)

          Value                              0.019
```

Figure ESI17. Mplus summary output for metric invariance (Step 2) with STEM majors data having modified I3 intercept highlighting fit information.

```
MODEL RESULTS                                   STDYX Standardization

                    Estimate                                        Estimate

Group STEM                                      Group STEM

 IC       BY
    I1                 1.000                       I1                 0.791
    I2                 1.034                       I2                 0.799
    I3                 1.025                       I3                 0.796
    I4                 1.053                       I4                 0.812

 CC       BY                                     CC       BY
    I5                 1.000                       I5                 0.788
    I6                 1.038                       I6                 0.807
    I7                 1.015                       I7                 0.796
    I8                 1.032                       I8                 0.803

 AC       BY                                     AC       BY
    I9                 1.000                       I9                 0.784
    I10                1.067                       I10                0.809
    I11                1.016                       I11                0.787
    I12                1.031                       I12                0.790

 CC       WITH                                   CC       WITH
    IC                 0.174                       IC                 0.295

 AC       WITH                                   AC       WITH
    IC                 0.119                       IC                 0.204
    CC                 0.085                       CC                 0.146

 Means                                          Means
    IC                 0.000                       IC                 0.000
    CC                 0.000                       CC                 0.000
    AC                 0.000                       AC                 0.000

 Intercepts                                     Intercepts
    I1                -0.004                       I1                -0.004
    I2                 0.012                       I2                 0.012
    I3                 1.974                       I3                 1.991
    I4                -0.003                       I4                -0.003
    I5                -0.009                       I5                -0.010
    I6                 0.004                       I6                 0.004
    I7                 0.033                       I7                 0.034
    I8                 0.013                       I8                 0.014
    I9                -0.018                       I9                -0.018
    I10               -0.006                       I10               -0.006
    I11               -0.001                       I11               -0.001
    I12               -0.021                       I12               -0.022

MODEL RESULTS                                   STDYX Standardization

                    Estimate                                        Estimate

Group NONSTEM                                   Group NONSTEM

 IC       BY
    I1                 1.000                       I1                 0.784
    I2                 1.034                       I2                 0.806
    I3                 1.025                       I3                 0.812
    I4                 1.053                       I4                 0.804

 CC       BY                                     CC       BY
    I5                 1.000                       I5                 0.801
    I6                 1.038                       I6                 0.819
    I7                 1.015                       I7                 0.791
    I8                 1.032                       I8                 0.802

 AC       BY                                     AC       BY
    I9                 1.000                       I9                 0.762
    I10                1.067                       I10                0.797
    I11                1.016                       I11                0.773
    I12                1.031                       I12                0.785

 CC       WITH                                   CC       WITH
    IC                 0.207                       IC                 0.335

 AC       WITH                                   AC       WITH
    IC                 0.145                       IC                 0.250
    CC                 0.104                       CC                 0.178

 Means                                          Means
    IC                 0.000                       IC                 0.000
    CC                 0.000                       CC                 0.000
    AC                 0.000                       AC                 0.000

 Intercepts                                     Intercepts
    I1                -0.008                       I1                -0.008
    I2                -0.019                       I2                -0.019
    I3                -0.028                       I3                -0.028
    I4                -0.050                       I4                -0.049
    I5                -0.023                       I5                -0.023
    I6                -0.053                       I6                -0.053
    I7                -0.036                       I7                -0.036
    I8                -0.005                       I8                -0.005
    I9                 0.041                       I9                 0.042
    I10                0.071                       I10                0.071
    I11                0.035                       I11                0.036
    I12                0.048                       I12                0.050
```

Figure ESI18. Mplus output for metric invariance (Step 2) with STEM majors data having modified I3 intercept highlighting unstandardized and standardized model parameters for both groups.

Step 3: Scalar Invariance (Strong)

Scalar invariance is tested by assigning the same parameter names to the intercept terms in both groups while also removing the restrictions on the mean of the factor terms for the second group using the * notation. As seen in Table 1 of the manuscript and in the R output, this significantly worsens the value of all fit indices (Figure ESI19) indicating that scalar invariance has not been achieved due to differences in loadings across groups. As before, the Mplus model parameters (Figure ESI20) are similar to those produced by R (Figure ESI12).

```
TITLE: Combined Dataset with Mean Differences Step 3 (Strong)
DATA: FILE = "InvarianceMean.dat";
VARIABLE:
NAMES = I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12 group;
USEVARIABLES ARE I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12;
GROUPING = group (1 = STEM 2 = NonSTEM);

MODEL:
! Model with standardized loading of first item on each factor
! Assigning a parameter name to each loading value (L1-12)
  IC BY I1@1 I2 I3 I4 (L1-L4);
  CC BY I5@1 I6 I7 I8 (L5-L8);
  AC BY I9@1 I10 I11 I12 (L9-L12);

! Setting factor intercepts to zero
  [IC@0];
  [CC@0];
  [AC@0];

! Allowing item intercepts to be freely estimated in one group
! assigning a parameter name so they will be equal across groups
  [I1-I12] (M1-M12);

! Allowing item error variances to be freely estimated
  I1-I12;

! Specifying the same model parameter names for the second group
! will cause loadings and item intercepts to be equivalent across
! groups while other parameters are freely estimated
MODEL NonSTEM:
  IC BY I1@1 I2 I3 I4 (L1-L4);
  CC BY I5@1 I6 I7 I8 (L5-L8);
  AC BY I9@1 I10 I11 I12 (L9-L12);

! Allowing factor intercepts vary
  [IC*];
  [CC*];
  [AC*];

  [I1-I12] (M1-M12);

  I1-I12;

OUTPUT:
STANDARDIZED;
```

```
MODEL FIT INFORMATION

Number of Free Parameters                       60

Loglikelihood

          H0 Value                       -28912.498
          H1 Value                       -27758.626

Information Criteria

          Akaike (AIC)                    57944.997
          Bayesian (BIC)                  58281.051
          Sample-Size Adjusted BIC        58090.428
            (n* = (n + 2) / 24)

Chi-Square Test of Model Fit

          Value                           2307.745
          Degrees of Freedom                   120
          P-Value                           0.0000

Chi-Square Contribution From Each Group

          STEM                             229.366
          NONSTEM                         2078.380

RMSEA (Root Mean Square Error Of Approximation)

          Estimate                           0.135
          90 Percent C.I.                    0.130  0.140
          Probability RMSEA <= .05           0.000

CFI/TLI

          CFI                                0.817
          TLI                                0.798

Chi-Square Test of Model Fit for the Baseline Model

          Value                          12068.162
          Degrees of Freedom                   132
          P-Value                           0.0000

SRMR (Standardized Root Mean Square Residual)

          Value                              0.238
```

Figure ESI19. Mplus summary output for scalar invariance (Step 3) with STEM majors data having modified I3 intercept highlighting fit information.

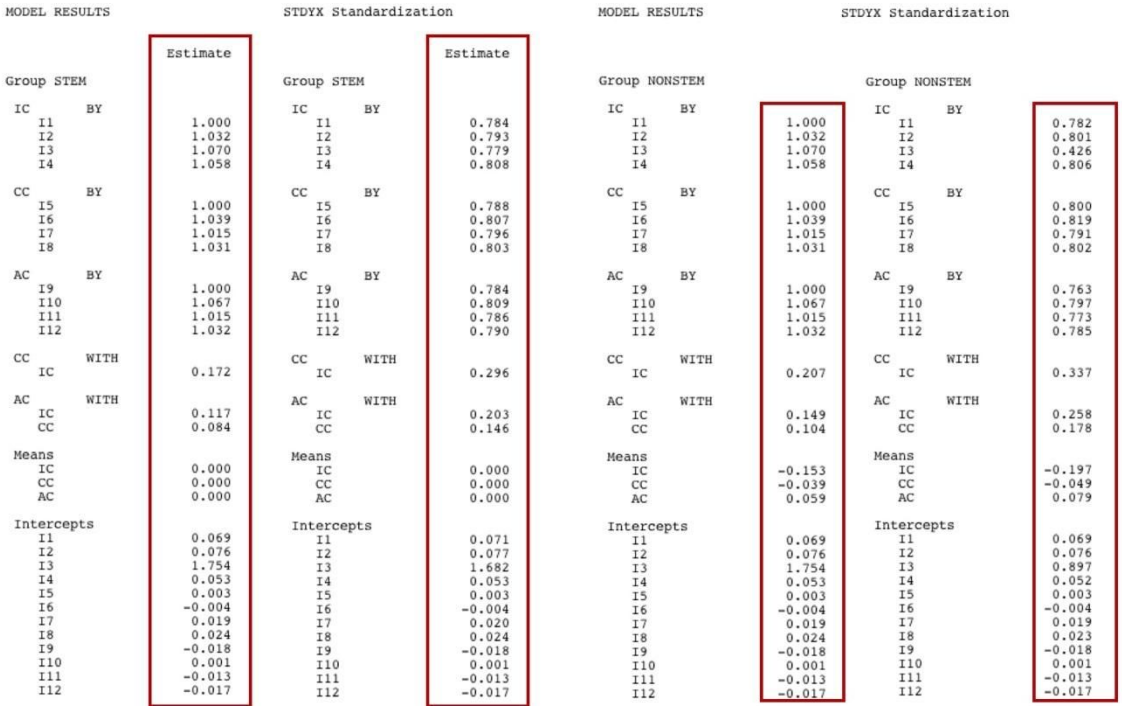


Figure ESI20. Mplus output for scalar invariance (Step 3) with STEM majors data having modified I3 intercept highlighting unstandardized and standardized model parameters for both groups.

Step 4: Conservative Invariance (Strict)

As noted previously, due to the poor fit of the scalar invariance model, you would stop at Step 3 and not go on to test Step 4 (conservative invariance with equal error variance terms). However, interested readers can test Step 4 in Mplus by providing the same name to the error variance parameters in both groups.

```
TITLE: Combined Dataset with Mean Differences Step 4 (Strict)
DATA: FILE = "InvarianceMean.dat";
VARIABLE:
NAMES = I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12 group;
USEVARIABLES ARE I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12;
GROUPING = group (1 = STEM 2 = NonSTEM);

MODEL:
! Model with standardized loading of first item on each factor
! Assigning a parameter name to each loading value (L1-12)
  IC BY I1@1 I2 I3 I4 (L1-L4);
  CC BY I5@1 I6 I7 I8 (L5-L8);
  AC BY I9@1 I10 I11 I12 (L9-L12);

! Setting factor intercepts to zero
  [IC@0];
  [CC@0];
  [AC@0];

! Allow item intercepts to be freely estimated in one group but
! assigning a parameter name so they will be equal across groups
  [I1-I12] (M1-M12);

! Allow item error variances to be freely estimated but
! assigning a parameter name so they will be equal across groups
  I1-I12 (E1-E12);

! Specifying the same model parameter names for the second group
! will cause loadings and item intercepts to be equivalent across
! groups while other parameters are freely estimated
MODEL NonSTEM:
  IC BY I1@1 I2 I3 I4 (L1-L4);
  CC BY I5@1 I6 I7 I8 (L5-L8);
  AC BY I9@1 I10 I11 I12 (L9-L12);

! Allowing factor intercepts vary
  [IC*];
  [CC*];
  [AC*];

  [I1-I12] (M1-M12);

  I1-I12(E1-E12);

OUTPUT:
STANDARDIZED;
```

Fit Indices for other Continuous Datasets

Tables ESI1 & ESI2 show the data-model fit output from R produced from following the previous steps with the two other continuous datasets: `combined` and `combined.invar.load`.

Table ESI1. Measurement Invariance Testing for the PRCQ Instrument Comparing STEM Majors and Non-STEM Majors With `combined` Simulated Data for Illustration

Step	Testing level	χ2	*df*	*p*-value	CFI	SRMR	RMSEA	Δχ2	Δ*df*	*p*-value	ΔCFI	ΔSRMR	ΔRMSEA
0	STEM majors Baseline	65	51	0.084	0.998	0.021	0.017	-	-	-	-	-	-
0	Non-STEM majors Baseline	52	51	0.437	1.000	0.016	0.004	-	-	-	-	-	-
1	Configural	117	102	0.142	0.999	0.018	0.012	-	-	-	-	-	-
2	Metric	120	111	0.245	0.999	0.019	0.009	3	9	0.964	0.000	0.001	0.003
3	Scalar	127	120	0.311	0.999	0.020	0.008	7	9	0.637	0.000	0.001	0.001
4	Conservative	135	132	0.417	1.000	0.020	0.005	8	12	0.786	0.001	0.000	0.003

Note. STEM majors $n = 1000$. Non-STEM majors $n = 1000$. Simulated data was used and altered at the scalar level (intercepts) for illustrative purposes; fit indices are from R.

Table ESI2. Measurement Invariance Testing for the PRCQ Instrument Comparing STEM Majors and Non-STEM Majors With `combined.invar.load` Simulated Data for Illustration

Step	Testing level	χ2	*df*	*p*-value	CFI	SRMR	RMSEA	Δχ2	Δ*df*	*p*-value	ΔCFI	ΔSRMR	ΔRMSEA
0	STEM majors Baseline	65	51	0.084	0.998	0.021	0.017	-	-	-	-	-	-
0	Non-STEM majors Baseline	66	51	0.081	0.997	0.017	0.017	-	-	-	-	-	-
1	Configural	131	102	0.028	0.997	0.019	0.017	-	-	-	-	-	-
2	Metric	305	111	< 0.001	0.983	0.051	0.042	101	9	< 0.001	0.014	0.032	0.025
3	Scalar	310	120	< 0.001	0.984	0.051	0.040	5	9	0.834	0.001	0.000	0.002
4	Conservative	433	132	< 0.001	0.974	0.043	0.048	123	12	< 0.001	0.010	0.008	0.008

Note. STEM majors $n = 1000$. Non-STEM majors $n = 1000$. Simulated data was used and altered at the scalar level (intercepts) for illustrative purposes; fit indices are from R.

Creating Ordered Categorical Data in R

As seen in the previous examples, the data simulation function in R creates continuous data which may not be representative of data collected from instruments used in chemistry education research, which often have five-point Likert-type scales. The code below is used to take the original simulated datasets and turn them into Likert-type data by collapsing the full ranges of data for each item into five bins using the `cut()` function. Note that this process of creating categorical data from continuous data ensures that each bin will be populated, but issues with testing models can arise if authentic categorical data are collected with empty bins (e.g., no responses in the 1 category).

```
STEM.ord<-STEM
for(i in 1:12){
  var[i]<-paste0("I", i)
  STEM.ord[[var[i]]]<-as.numeric(cut(STEM[[var[i]]], breaks=5))
                }

nonSTEM.ord<-nonSTEM
for(i in 1:12){
  var[i]<-paste0("I", i)
  nonSTEM.ord[[var[i]]]<-as.numeric(cut(nonSTEM[[var[i]]],
breaks=5))
                }
combined.ord<-rbind(STEM.ord, nonSTEM.ord)


nonSTEM.invar.load.ord<-nonSTEM.invar.load
for(i in 1:12){
  var[i]<-paste0("I", i)
  nonSTEM.invar.load.ord[[var[i]]]<-
as.numeric(cut(nonSTEM.invar.load[[var[i]]], breaks=5))
}
combined.invar.load.ord<-rbind(STEM.ord, nonSTEM.invar.load.ord)

STEM.invar.mean.ord<-STEM.invar.mean
for(i in 1:12){
  var[i]<-paste0("I", i)
  STEM.invar.mean.ord[[var[i]]]<-
as.numeric(cut(STEM.invar.mean[[var[i]]], breaks=5))
}
combined.invar.mean.ord<-rbind(STEM.invar.mean.ord, nonSTEM.ord)
```

When data collected on Likert-type scales have fewer than seven categories or the full range of the response scale is not used by most respondents (i.e. a ceiling or floor effect) it is often recommended to treat the data as ordinal categorical data rather than continuous. In a factor analysis framework, this type of data is best modeled using a robust diagonally weighted least squares estimator, such as WLSMV (Finney and DiStefano, 2013). A noticeable difference in working with ordinal data the software will compute thresholds which are used to map the categorical variables onto an assumed underlying normal distribution of latent item responses and therefore create a set of latent correlations. This process is can be conceptualized as the

reverse of the process used to create ordered categorical data from the original continuous data show in prior steps.

The concept of thresholds can be visualized by plotting the distribution of values for an item both in its continuous and categorical form. For this example, responses to I1 in the continuous data are visualized with a density plot (Figure ESI21a) and I1 responses in the categorical data are visualized with a bar plot (Figure ESI21b) using the code below.

```
plot(density(combined$I1),
     main="Density Plot for Combined Data Item I1 - Continuous",
     ylab="Frequency", xlab="Response")

barplot(prop.table(table(combined.ord$I1)),
        main="Frequency Plot for Combined Data Item I1 - Ordinal",
        ylab="Frequency", xlab="Response")
```

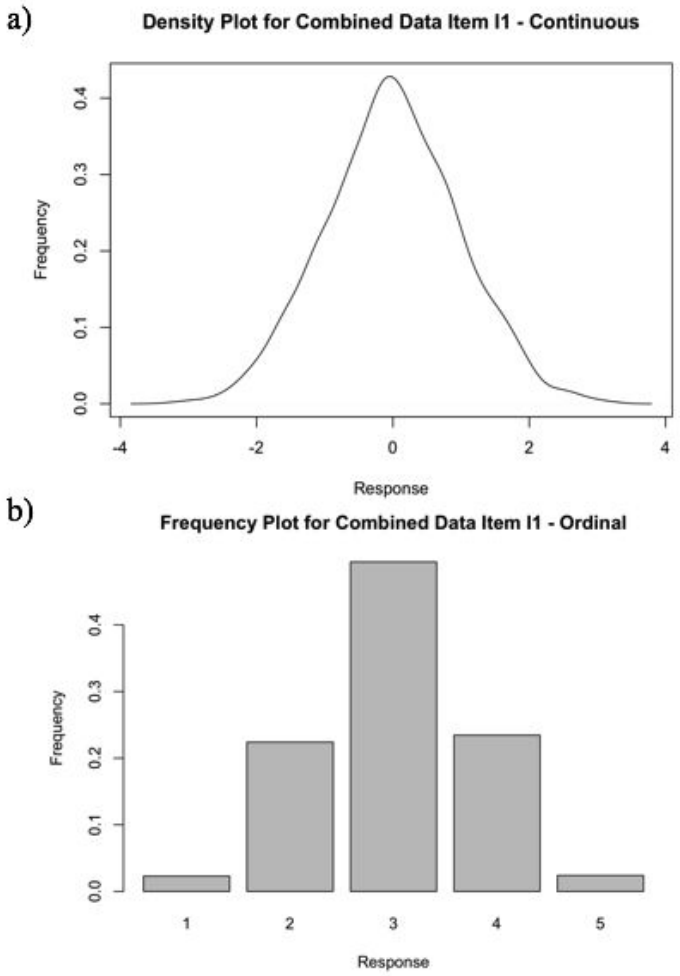


Figure ESI21. Density plot of continuous I1 responses (a) and frequency plot of categorical I1 responses (b)

Visual inspection of the two plots shows how the original continuous distribution aligns with the categorical data in that the middle responses have higher response frequencies and the extreme responses have lower response frequencies. When the ordinal data in Figure ESI21b are

used to estimate a factor model, the software will assume the categorical data are representative of an underlying continuous variable (DiStefano and Morgan, 2014) and determine cut points, called thresholds, where the unobserved continuous distribution would have been divided to create the observed categorical distribution.

Since the categorical data used in this example were created from continuous data, we are able find the true cut points using the same code as before.

```
summary(cut(combined$I1, breaks=5))
```

Plotting these cut points (–1.97, –0.672, 0.624, and 1.92) on the continuous distribution (Figure ESI22) shows how the categorical data were simulated, and also provides insight into how the factor analysis itself will identify thresholds in the categorical data.

```
plot(density(combined$I1), main="Density Plot for Combined Data
     Item I1 - Continuous",ylab="Frequency", xlab="Response")
abline(v=c(-1.97, -0.672, 0.624, 1.92), col="grey")
```

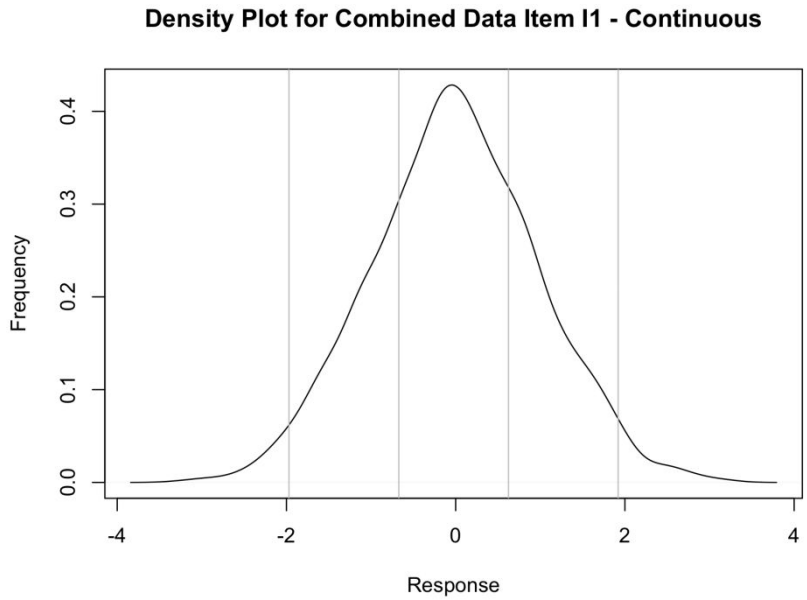


Figure ESI22. Density plot of continuous I1 responses showing cut points used to create categorical data.

Estimating Models with Ordered Categorical Data in R and Mplus

Running the factor models in R and also exporting the data for running in Mplus will provide an opportunity to see the threshold values established by the software. Full measurement invariance testing steps will be described in later sections. Both programs will automatically switch to the correct estimator (WLSMV) when informed that the data are not continuous. In lavaan syntax the argument `ordered` is used.

```
combined.ord.cfa<-cfa(data = combined.ord, model = model.test,
                      ordered=c("I1", "I2", "I3", "I4", "I5",
                      "I6", "I7", "I8", "I9", "I10", "I11",
                      "I12"))
summary(combined.ord.cfa, standardized=TRUE, fit.measures=TRUE)

combined.ord$group<-combined.ord$group %>% as.factor() %>%
                    as.numeric()
prepareMplusData(combined.ord, filename="CombinedOrdinal.dat",
                 inpfile = T, keepCols=c("I1", "I2", "I3",
                 "I4","I5", "I6", "I7", "I8", "I9", "I10", "I11",
                 "I12", "group"))
```

In Mplus the variables are specified as `categorical`.

```
TITLE: Combined Ordinal Data - CFA Model
DATA: FILE = "CombinedOrdinal.dat";
VARIABLE:
NAMES = I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12 group;
MISSING=.;

USEVARIABLES ARE I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12;
CATEGORICAL ARE I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12;

MODEL:
IC BY I1 I2 I3 I4;
CC BY I5 I6 I7 I8;
AC BY I9 I10 I11 I12;

OUTPUT:
STANDARDIZED;
```

The full output of both programs can be examined to confirm similarities in how the data are treated as well as the matched fit indices and model parameters. Figure ESI23 shows the threshold values calculated by each program, indicated with the `t` notation in R and the `$` notation in Mplus. As expected, the thresholds for I1 are similar to those used to create the categorical data from the continuous, even though neither R or Mplus had access to the continuous data when generating the threshold values.

a)

Thresholds:

	Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
I1|t1	-1.919	0.058	-33.205	0.000	-1.919	-1.919
I1|t2	-0.645	0.030	-21.314	0.000	-0.645	-0.645
I1|t3	0.674	0.030	22.131	0.000	0.674	0.674
I1|t4	1.927	0.058	33.130	0.000	1.927	1.927

b)

Thresholds

I1$1	-1.919	0.058	-33.213	0.000
I1$2	-0.645	0.030	-21.319	0.000
I1$3	0.674	0.030	22.137	0.000
I1$4	1.927	0.058	33.138	0.000

Figure ESI23. Threshold values from R (a) and Mplus (b)

Data Model Fit for Ordered Categorical Data with WLSMV Estimator

The fit index cut off values recommended by Hu and Bentler (1999) were based on work using the maximum likelihood (ML) estimator which is appropriate for continuous data. Since a different estimator is used with categorical data, it is not appropriate to use the same Hu and Bentler recommendations for fit index cut off values. Simulation studies with the WLSMV estimator have indicated that more rigorous cut off values are best, particularly when the data contain a small number of categories or are severely nonnormal (Yu, 2002; Beauducel and Herzberg, 2006; DiStefano and Morgan, 2014). Recommendations for fit index values with the WLSMV estimator are CFI $\geq$ 0.95 and RMSEA $\leq$ 0.05. The SRMR is not recommended with the WLSMV estimator. In the context of invariance testing, less work has been done to determine recommended values for change in CFI and RMSEA values between models compared to the ML estimator. As with the fit indices themselves, simulation studies suggest either using more rigorous ΔCFI and ΔRMSEA values than those used with ML estimation or providing multiple sources of justification for acceptable data-model fit potentially using different estimators to see if similar conclusions about invariance would be drawn (Sass *et al.*, 2014).

Invariance Testing with R – Ordered Categorical Data

Measurement invariance testing in R with categorical data can be conducted following similar steps as those used for continuous data. However, it should be noted that other researchers have advocated for a different order of steps or different sets of constraints when working with categorical data (Millsap and Yun-Tein, 2004; Wu and Estabrook, 2016; Svetina *et al.*, 2019). The primary differences when working with categorical data compared to continuous are that the ordinal nature of the data must be specified in order for the correct estimator to be used, and thresholds must be constrained along with other model parameters during invariance testing steps.

Also, unique to working with categorical data, a decision must be made about scaling of the underlying latent normal distribution for each set of item responses using either delta or theta scaling. In delta scaling the total variance of the latent response is set to 1 and in theta scaling the variance of the residual term is set to 1. These decisions primarily influence how the model parameters are identified. Theta scaling is appropriate for invariance research (Millsap and Yun-Tein, 2004) and was chosen for the analysis here, but it is possible to convert parameters between delta and theta scaling (Finney and DiStefano, 2013). Since theta scaling affects the residual terms, Step 4 of invariance testing (strict) is not necessary with categorical data when following this method.

The steps taken in this ESI will parallel those used previously for continuous data. The data used in this section are the categorical version of the continuous data used in previous examples where the mean for I3 was changed in the STEM majors group. The code for all steps of invariance testing in R with categorical data are specified below and the fit statistics are summarized in Table ESI3 using the WLSMV output from lavaan as given in the `Robust` column. Fit statistics for models using the other categorical datasets are provided in Tables ESI4 & ESI5.

Step 0: Establishing Baseline Model

The baseline model for each group is specified in the same way as the continuous data but now using the ordinal data set and specifying which variables are ordered categorical as well as the use of the theta parameterization. The same three factor model used for the continuous data is used for the categorical data.

```
STEM.step0.ord<-cfa(data = combined.invar.mean.ord %>%
                    filter(group==STEM), model=model.test,
                    ordered=c("I1", "I2", "I3", "I4", "I5", "I6",
                    "I7", "I8", "I9", "I10", "I11", "I12"),
                    parameterization="theta")

summary(STEM.step0.ord, standardized=TRUE, fit.measures=TRUE)

nonSTEM.step0.ord<-cfa(data=combined.invar.mean.ord %>%
                       filter(group=="nonSTEM"),
                       model=model.test, ordered=c("I1", "I2",
                       "I3", "I4", "I5", "I6", "I7", "I8", "I9",
                       "I10", "I11", "I12"),
                       parameterization="theta")

summary(nonSTEM.step0.ord, standardized=TRUE, fit.measures=TRUE)
```

Step 1: Configural Invariance

Configural invariance uses data from both groups while specifying the grouping variable.

```
step1.comb.mean.ord<-cfa(data=combined.invar.mean.ord,
                         group="group", model=model.test,
                         ordered=c("I1", "I2", "I3", "I4", "I5",
                         "I6", "I7", "I8", "I9", "I10", "I11",
                         "I12"), parameterization="theta")

 summary(step1.comb.mean.ord, standardized=TRUE,
         fit.measures=TRUE)
```

Step 2: Metric Invariance (Weak)

Metric invariance is tested by holding the loadings equal across groups.

```
step2.comb.mean.ord<-cfa(data=combined.invar.mean.ord,
                         group="group", model=model.test,
                         ordered=c("I1", "I2", "I3", "I4", "I5",
                         "I6", "I7", "I8", "I9", "I10", "I11",
                         "I12"), group.equal=c("loadings"),
                         parameterization="theta")

summary(step2.comb.mean.ord, standardized=TRUE,
fit.measures=TRUE)
```

Step 3: Scalar Invariance (Strong)

Adding the constraint of equal thresholds across groups is similar to holding intercepts equal to test for scalar invariance in continuous data.

```
step3.comb.mean.ord<-cfa(data=combined.invar.mean.ord,
                         group="group", model=model.test,
                         ordered=c("I1", "I2", "I3", "I4", "I5",
                         "I6", "I7", "I8", "I9", "I10", "I11",
                         "I12"), group.equal=c("loadings",
                         "thresholds"), parameterization="theta")

summary(step3.comb.mean.ord, standardized=TRUE,
fit.measures=TRUE)
```

Table ESI3. Measurement Invariance Testing for the PRCQ Instrument Comparing STEM Majors and Non-STEM Majors With `combined.invar.mean` Simulated Categorical Data for Illustration

Step	Testing level	χ2	*df*	*p*-value	CFI	RMSEA	Δχ2	Δ*df*	*p*-value	ΔCFI	ΔRMSEA
0	STEM majors Baseline	81	51	0.005	0.996	0.024	-	-	-	-	-
0	Non-STEM majors Baseline	61	51	0.162	0.999	0.014	-	-	-	-	-
1	Configural	142	102	0.006	0.997	0.020	-	-	-	-	-
2	Metric	145	111	0.017	0.998	0.018	3	9	0.231	0.001	0.002
3	Scalar	869	144	< 0.001	0.953	0.071	724	9	< 0.001	0.045	0.053

Note. STEM majors $n = 1000$. Non-STEM majors $n = 1000$. Simulated data was used and altered at the scalar level (intercepts) for illustrative purposes; fit indices are from R.

Invariance Testing with Mplus – Ordered Categorical Data

Following the previously shown steps, the categorical data in R are exported to Mplus by first converting the group variable from a text format into a numeric format.

```
combined.invar.mean.ord$group<-combined.ord$group %>% as.factor()
                                %>% as.numeric()

prepareMplusData(combined.invar.mean.ord,
                 filename="CombinedInvarMeanOrdinal.dat",
                 inpfile = T, keepCols=c("I1", "I2", "I3",
                 "I4","I5", "I6", "I7", "I8", "I9", "I10",
                 "I11", "I12", "group"))
```

As with lavaan, the default estimator in Mplus is ML but the software will adjust to an appropriate estimator for ordinal data (WLSMV) by specifying the item variables as categorical. The call for theta parameterization is also added and the models are specified separately for each group. Following these steps for R and Mplus should provide similar fit indices and model parameters.

Step 0: Establishing Baseline Model

```
TITLE: Categorical STEM Majors Group Step 0
DATA: FILE = "CombinedInvarMeanOrdinal.dat";
VARIABLE:
NAMES = I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12 group;
MISSING=.;

USEVARIABLES ARE I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12;
CATEGORICAL ARE I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12;
USEOBSERVATIONS are group==2;

ANALYSIS: PARAMETERIZATION=THETA;

MODEL:
IC BY I1 I2 I3 I4;
CC BY I5 I6 I7 I8;
AC BY I9 I10 I11 I12;

OUTPUT:
STANDARDIZED;


TITLE: Categorical Non-STEM Majors Group Step 0
DATA: FILE = "CombinedInvarMeanOrdinal.dat";
VARIABLE:
NAMES = I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12 group;
MISSING=.;

USEVARIABLES ARE I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12;
CATEGORICAL ARE I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12;
USEOBSERVATIONS are group==1;

ANALYSIS: PARAMETERIZATION=THETA;

MODEL:
IC BY I1 I2 I3 I4;
CC BY I5 I6 I7 I8;
AC BY I9 I10 I11 I12;

OUTPUT:
STANDARDIZED;
```

Step 1: Configural Invariance

By default, Mplus will constrain thresholds equal across groups so this must be released by freeing all thresholds for all variables. The notation to free the thresholds uses the $ character. Four thresholds must be freed since four thresholds would be required to divide the underlying continuous distribution into five categories. As was done with the continuous data, the factor means are set to zero. The error variances are set to one for categorical data, in line with theta parameterization.

```
TITLE: Categorical Combined Dataset with Mean Differences Step 1
(Configural)
DATA: FILE = "CombinedInvarMeanOrdinal.dat";
VARIABLE:
NAMES = I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12 group;
CATEGORICAL ARE I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12;
USEVARIABLES ARE I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12;
GROUPING = group (1 = NonSTEM 2 = STEM);

ANALYSIS: PARAMETERIZATION=THETA;

MODEL:
! Model with standardized loading of first item on each factor
  IC BY I1@1 I2 I3 I4;
  CC BY I5@1 I6 I7 I8;
  AC BY I9@1 I10 I11 I12;

! Freeing Thresholds
  [I1$1-I12$1*];
  [I1$2-I12$2*];
  [I1$3-I12$3*];
  [I1$4-I12$4*];

! Set factor means to 0
  [IC@0];
  [CC@0];
  [AC@0];

! Set error variances to 1
  I1-I12@1

MODEL STEM:
  IC BY I1@1 I2 I3 I4;
  CC BY I5@1 I6 I7 I8;
  AC BY I9@1 I10 I11 I12;

! Freeing Thresholds
  [I1$1-I12$1*];
  [I1$2-I12$2*];
  [I1$3-I12$3*];
  [I1$4-I12$4*];
```

```
! Set factor means to 0
  [IC@0];
  [CC@0];
  [AC@0];

! Set error variances to 1
  I1-I12@1

OUTPUT:
STANDARDIZED;
```

Step 2: Metric Invariance (Weak)

Loadings are constrained equal across groups by assigning the same name to the parameters in both groups. This is the same method used for invariance testing with the continuous data.

```
TITLE: Categorical Combined Dataset with Mean Differences Step 2
DATA: FILE = "CombinedInvarMeanOrdinal.dat";
VARIABLE:
NAMES = I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12 group;
CATEGORICAL ARE I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12;
USEVARIABLES ARE I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12;
GROUPING = group (1 = NonSTEM 2 = STEM);

ANALYSIS: PARAMETERIZATION=THETA;

MODEL:
! Model with standardized loading of first item on each factor
! Assigning a parameter name to each loading value (L1-L12)
  IC BY I1@1 I2 I3 I4 (L1-L4);
  CC BY I5@1 I6 I7 I8 (L5-L8);
  AC BY I9@1 I10 I11 I12 (L9-L12);

! Freeing Thresholds
  [I1$1-I12$1*];
  [I1$2-I12$2*];
  [I1$3-I12$3*];
  [I1$4-I12$4*];

! Set factor means to 0
  [IC@0];
  [CC@0];
  [AC@0];

! Set error variances to 1
  I1-I12@1

MODEL STEM:
  IC BY I1@1 I2 I3 I4 (L1-L4);
  CC BY I5@1 I6 I7 I8 (L5-L8);
  AC BY I9@1 I10 I11 I12 (L9-L12);
```

```
  ! Freeing Thresholds
  [I1$1-I12$1*];
  [I1$2-I12$2*];
  [I1$3-I12$3*];
  [I1$4-I12$4*];

  ! Set factor means to 0
  [IC@0];
  [CC@0];
  [AC@0];

 ! Set error variances to 1
  I1-I12@1

OUTPUT:
STANDARDIZED;
```

Step 3: Scalar Invariance (Strong)

Mplus and lavaan differ in their default settings when thresholds are constrained equal across groups. To mimic the lavaan output the factor means and error variance terms for the second group are freed in the Mplus code. Freeing these parameters also aligns scalar invariance testing in the categorical data with the same step for the continuous data. Recall that the goal of Step 3 is to determine if the factors are being measured on the same scale in each group so that factor means can be compared across groups. Therefore, one group should have a mean of zero in order to function as a reference while the mean of the other group is freely estimated.

```
TITLE: Categorical Combined Dataset with Mean Differences Step 3
DATA: FILE = "CombinedInvarMeanOrdinal.dat";
VARIABLE:
NAMES = I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12 group;
CATEGORICAL ARE I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12;
USEVARIABLES ARE I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12;
GROUPING = group (1 = NonSTEM 2 = STEM);

ANALYSIS: PARAMETERIZATION=THETA;

MODEL:
  IC BY I1@1 I2 I3 I4 (L1-L4);
  CC BY I5@1 I6 I7 I8 (L5-L8);
  AC BY I9@1 I10 I11 I12 (L9-L12);

  [I1$1-I12$1*];
  [I1$2-I12$2*];
  [I1$3-I12$3*];
  [I1$4-I12$4*];

  [IC@0];
  [CC@0];
  [AC@0];

  I1-I12@1
```

```
MODEL STEM:
  IC BY I1@1 I2 I3 I4 (L1-L4);
  CC BY I5@1 I6 I7 I8 (L5-L8);
  AC BY I9@1 I10 I11 I12 (L9-L12);

! Fix thresholds equal by not specifying for this group

! Set factor means free
[IC*];
[CC*];
[AC*];

! Set error variances free
I1-I12*

OUTPUT:
STANDARDIZED;
```

Fit Indices for Invariance Testing Steps with other Simulated Categorical Data

Tables ESI4 & 5 show the data-model fit output from R produced from following the previous steps with the two other categorical datasets: `combined.ord` and `combined.invar.load.ord`.

Table ESI4. Measurement Invariance Testing for the PRCQ Instrument Comparing STEM Majors and Non-STEM Majors With `combined.ord` Simulated Categorical Data for Illustration

Step	Testing level	$\chi2$	*df*	*p*-value	CFI	RMSEA	$\Delta\chi2$	Δdf	*p*-value	ΔCFI	ΔRMSEA
0	STEM majors Baseline	81	51	0.005	0.996	0.024	-	-	-	-	-
0	Non-STEM majors Baseline	61	51	0.162	0.999	0.014	-	-	-	-	-
1	Configural	142	102	0.006	0.997	0.020	-	-	-	-	-
2	Metric	145	111	0.017	0.998	0.018	3	9	0.964	0.001	0.002
3	Scalar	869	144	<0.001	0.953	0.071	724	33	< 0.001	0.045	0.053

Note. STEM majors $n = 1000$. Non-STEM majors $n = 1000$. Simulated data was used and altered at the scalar level (intercepts) for illustrative purposes; fit indices are from R.

Table ESI5. Measurement Invariance Testing for the PRCQ Instrument Comparing STEM Majors and Non-STEM Majors With `combined.invar.load.ord` Simulated Categorical Data for Illustration

Step	Testing level	χ2	*df*	*p*-value	CFI	RMSEA	Δχ2	Δ*df*	*p*-value	ΔCFI	ΔRMSEA
0	STEM majors Baseline	81	51	0.005	0.996	0.024	-	-	-	-	-
0	Non-STEM majors Baseline	40	51	0.869	1.000	0.000	-	-	-	-	-
1	Configural	119	102	0.120	0.999	0.013	-	-	-	-	-
2	Metric	383	111	< 0.001	0.982	0.050	264	9	< 0.001	0.017	0.037
3	Scalar	1305	144	< 0.001	0.925	0.090	922	33	< 0.001	0.057	0.040

Note. STEM majors $n = 1000$. Non-STEM majors $n = 1000$. Simulated data was used and altered at the scalar level (intercepts) for illustrative purposes; fit indices are from R.

ESI References

Beauducel A. and Herzberg P. Y., (2006), On the Performance of Maximum Likelihood Versus Means and Variance Adjusted Weighted Least Squares Estimation in CFA. *Struct. Equ. Model. A Multidiscip. J.*, **13**(2), 186–203.

Bontempo D. E. and Hofer S. M., (2007), Assessing Factorial Invariance in Cross-Sectional and Longitudinal Studies., in Ong A. D. and van Dulmen M. H. M. (eds.), *Series in positive psychology. oxford handbook of methods in positive psychology*. Oxford University Press, pp. 153–175.

Byrne B. M., (2004), Testing for multigroup invariance using AMOS Graphics: A road less traveled. *Struct. Equ. Model. A Multidiscip. J.*, **11**(2), 272–300.

DiStefano C. and Morgan G. B., (2014), A Comparison of Diagonal Weighted Least Squares Robust Estimation Techniques for Ordinal Data. *Struct. Equ. Model.*, **21**, 425–438.

Finney S. J. and DiStefano C., (2013), Non-normal and categorical data in structural equation modeling., in Hancock G. R. and Mueller R. O. (eds.), *Structural equation modeling: a second course*. Charlotte, NC: Information Age Publishing, pp. 439–492.

Hallquist M. N. and Wiley J. F., (2018), MplusAutomation: An R Package for Facilitating Large-Scale Latent Variable Analyses in Mplus. *Struct. Equ. Model.*, **25**(4), 621–638.

Hancock G. R., Stapleton L. M., and Arnold-Berkovits I., (2009), The tenuousness of invariance tests within multisample covariance and mean structure models., in *Structural equation modeling in educational research: concepts and applications*., pp. 137–174.

Hirschfeld G. and Von Brachel R., (2014), Multiple-Group confirmatory factor analysis in R – A tutorial in measurement invariance with continuous and ordinal. *Pract. Assessment, Res. Eval.*, **19**(7), 1–11.

Hu L. and Bentler P. M., (1999), Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Struct. Equ. Model. A Multidiscip. J.*, **6**(1), 1–55.

Komperda R., (2017), Likert-Type Survey Data Analysis with R and RStudio., in Gupta T. (ed.), *Computer-aided data analysis in chemical education research (cadacer): advances and avenues*. Washington, DC: ACS Symposium Series; American Chemical Society, pp. 91–116.

Millsap R. E. and Yun-Tein J., (2004), Assessing Factorial Invariance in Ordered-Categorical Measures. *Multivariate Behav. Res.*, **39**(3), 479–515.

Muthén L. K. and Muthén B. O., (2017), Mplus User's Guide, Eighth. Los Angeles, CA: Muthén & Muthén.

Narayanan A., (2012), A review of eight software packages for structural equation modeling. *Am. Stat.*, **66**(2), 129–138.

R Core Team, (2019), *R: A language and environment for statistical computing*, [Computer software].

Revelle W., (2018), *psych: procedures for psychological, psychometric, and personality research*, [Computer software].

Rosseel Y., (2020), The lavaan Project. http://lavaan.ugent.be/

Rosseel Y., (2012), lavaan: An R Package for Structural Equation Modeling. *J. Stat. Softw.*, **48**(2), 1–36.

Sass D. A., Schmitt T. A., and Marsh H. W., (2014), Evaluating Model Fit With Ordered Categorical Data Within a Measurement Invariance Framework: A Comparison of Estimators. *Struct. Equ. Model.*, **21**(2), 167–180.

Schneider W. J., (2019), *simstandard: generate standardized data*, [Computer software].

Svetina D., Rutkowski L., and Rutkowski D., (2019), Multiple-Group Invariance with Categorical Outcomes Using Updated Guidelines: An Illustration Using Mplus and the lavaan/semTools Packages. *Struct. Equ. Model.*, **0**(0), 1–20.

Wei T. and Simko V., (2017), *R package "corrplot": visualization of a correlation matrix,* [Computer software].

Wickham H., (2007), Reshaping Data with the reshape Package. *J. Stat. Softw.*, **21**(12), 1–20.

Wickham H., (2016), ggplot2: Elegant Graphics for Data Analysis, New York: Springer-Verlag.

Wickham H., François R., Henry L., and Müller K., (2019), *dplyr: a grammar of data manipulation,* [Computer software].

Wu H. and Estabrook R., (2016), Identification of Confirmatory Factor Analysis Models of Different Levels of Invariance for Ordered Categorical Outcomes. *Psychometrika*, **81**(4), 1014–1045.

Yu C.-Y., (2002), Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes.