Multigroup Invariance Testing for Cross-Cultural Research

Johannes A. Karl

Contents

Abstract

Cross-cultural research is essential for understanding human behavior across diverse societies, but ensuring measurement equivalence across cultures is challenging. This chapter introduces multigroup invariance testing as a crucial method for addressing this challenge. It explains the concept of measurement invariance and its role in valid cross-cultural comparisons. The chapter outlines the theoretical foundations and statistical principles of multigroup invariance testing. The chapter illustrates how this technique assesses the equivalence of measurement tools, like surveys, across different cultural groups. It covers different levels of invariance, including configural, metric, and scalar invariance, enabling robust cross-cultural comparisons. The chapter emphasizes how multigroup invariance testing enhances cross-cultural research validity by disentangling true cultural differences from measurement bias. It provides

J. A. Karl (✉)
School of Psychology, Victoria University of Wellington, Wellington, New Zealand

School of Psychology, Dublin City University, Dublin, Ireland
e-mail: johannes.karl@dcu.ie

C. U. Krägeloh et al. (eds.), *International Handbook of Behavioral Health Assessment*,
https://doi.org/10.1007/978-3-030-89738-3_7-1

guidance on avoiding common pitfalls and offers practical recommendations for effective implementation. This chapter aims to serve as a guide for researchers and practitioners in cross-cultural studies and health studies. By understanding and applying multigroup invariance testing, they can conduct rigorous and insightful research that captures the nuances of human behavior across diverse cultural contexts.

Keywords

Equivalence · Invariance · Multigroup · Confirmatory factor analysis · Cross-cultural

Introduction

Humanity lives in a global world, with an increasing number of researchers becoming aware that the hegemony of Western constructs in psychological science should not be confused with evidence of their universality, leading to a rise of interest in cross-cultural research (Boer et al., 2018). The overwhelming majority of research on behavioral health has been conducted by Western (specifically US) researchers on Western samples (Henrich, 2020). This fact should by no means be taken as an indication that this previous research did and does not have value but should rather highlight the limited scope both conceptually and psychometrically that has dominated psychological science over the past decades. This chapter aims to draw attention to challenges and opportunities for behavioral health research in cross-cultural research and cross-group research. Specifically, it will focus on the theoretical and practical issues underpinning one of the basic requirements of cross-group research, *multigroup measurement equivalence*.

The core of the problem posed by comparing constructs across groups is that most constructs in behavioral health research are not directly accessible, but rather need to be inferred from a range of indicators. Taking, for example, constructs such as anxiety (Kroenke et al., 2007) and depression (Kroenke & Spitzer, 2002), these need to be inferred from responses to a range of items that are supposed to measure the underlying construct. These are commonly referred to as latent and observed variables, with unobservable latent variables being expressed in observable variables such as recordings of sleep disturbance. There is a good epistemological debate to be had about when one should consider a variable to be observed (Borsboom, 2008) and if one should pin the latent status of a variable to its unobservability or as Bollen (2002) puts it: "[This definition of a variable as latent] presupposes knowledge that it will never be possible to directly measure these variables. Certainly, we do not now have the technology or knowledge to do so, but we cannot say that it will never be possible" (p. 614). While the exact definition of what constitutes a latent variable likely depends on context and the specific statistical model used by the researcher, Bollen (2002) has proposed a broad definition of a latent variable subsuming other more specific definitions, such as local independence as: "A latent random

(or non-random) variable is a random (or non-random) variable for which there is no sample realization for at least some observations in a given sample" (p. 612).

This broad definition (encompassing all variables that one cannot directly observe and therefore lack a sample realization, or that one failed to measure and therefore have missing data) lets us get to one of the core issues of cross-cultural research, which is the estimation of these absent sample realizations (or latent variable) in different groups and the equivalence of these estimations. This issue has long been recognized not only in cross-cultural research (van de Vijver & Leung, 1997; Vandenberg & Lance, 2000), but also in research focusing on subgroups such as gender (Van Doren et al., 2021; Zager Kocjan et al., 2021), sex (Waldren et al., 2022), and even groups such as parents (Wang et al., 2006). Nevertheless, while the problem of equivalence is known, it is commonly disregarded in applied research, with some studies showing that only 4% of published papers in a selected sample investigated group comparability (D'Urso et al., 2022). This neglect to establish proper group comparability can have drastic consequences for the inference drawn from multigroup investigations, leading to wrong inferences about group differences (Christopher et al., 2009; Jeong & Lee, 2019; Karl et al., 2020; Wu & Huang, 2014). This chapter aims to support research in behavioral health by highlighting the conceptual and methodological underpinnings of establishing *invariance and equivalence* in a multigroup research setting.

Core to psychometric applications across cultures and socio-demographic subgroups is the concept of bias and equivalence. These can be broken down broadly into four levels, which internally can be subdivided into a number of aspects. As it is easier to work through classifications with a pragmatic example, it will be beneficial to lay out the ones used in this chapter.

Conceptual Equivalence

Let us engage in a small thought experiment. Imagine you are in a room with a robot. Sadly, there was a manufacturing mistake, and the robot was not fitted with a module to experience enjoyment. You are both given a chocolate chip cookie and are asked about your current level of enjoyment along several dimensions, taste, smell, and appearance. You might be able to provide an answer, but your robot companion will not be able to produce a comparable answer as the concept is fully alien to them.

What would a lack of *conceptual equivalence* mean in practical terms for a researcher? Essentially, it would preclude any meaningful comparison between cultures, and a researcher should be quite excited about this finding. Establishing the lack of conceptual equivalence of a construct that was initially considered to be universal opens up a potentially rich research area as past research on culture-bound syndromes has shown (Simons & Hughes, 1985; Tseng, 2006). Importantly, issues of conceptual non-equivalence mostly arise at the pre-statistical stage during scale translations and consultations with cultural experts.

Functional Equivalence

To continue our thought example, a technician has entered the room and installed the missing enjoyment module, but in their haste, they wired it up wrong and connected it to the robot's pain sensors. You are given another cookie and get asked about your current level of enjoyment based on taste, smell, and appearance. You both now respond, but your answers differ radically. While you enjoyed the cookie, the robot returns no enjoyment after querying the joy module. You both now have a concept you label as enjoyment, but it fulfills functionally very different roles.

While some researchers have indicated that functional equivalence is present if in different cultures people show the same behavior to meet a functional end in the same situation (such as displaying help-seeking behavior if distressed to find support). This limits potential empirical tests to behavioral observations and behaviors that can be observed (Hui & Triandis, 1985). When working with more abstract concepts, this approach is often not viable, and researchers have relied on conceptual nomological networks. *Nomological networks* refer to the complex web of relationships between psychological constructs, such as traits, behaviors, and attitudes. These networks are often studied by ethnographers and anthropologists to gain a deeper understanding of cultural practices and belief systems by examining the functional role behaviors play in a wider network of practices. However, they have historically been overlooked by psychologists, who have instead focused on the relationships between individual constructs and their predictive power. Recently, there has been a renewed interest in nomological networks among psychologists, particularly with the rise of network methods in the field. Large-scale projects, such as the validation of the revised moral foundations questionnaire, have explicitly employed nomological networks in cross-cultural research, highlighting their importance in understanding complex psychological phenomena across different cultures by defining theoretical relationships of expected relationships between psychological constructs (Atari et al., 2023).

Nomological networks are one core method to systematically map *out functional equivalence* between groups by helping researchers to identify if a behavior is utilized in the same context across groups. One of the main reasons for this increased interest in nomological networks is the availability of new network methods in psychology. These methods allow researchers to conduct substantive analyses on nomological networks, including comparison approaches to Ising models and aggregated meta-networks. These techniques enable researchers to identify important network structures, such as central nodes and communities, and explore the relationships between different constructs in the network (Borsboom et al., 2021). Overall, the recent rise in interest in nomological networks reflects a growing recognition of the importance of understanding the complex relationships between psychological constructs and the potential insights that can be gained through network analysis. As this method continues to evolve and becomes more sophisticated, it is likely that nomological networks will play an increasingly important role in psychological

research. Nevertheless, it is important to consider that nomological networks might not be free of bias (Fischer et al., 2022).

Structural Equivalence

To continue our thought example, again a technician has entered the room and correctly wires the joy module to olfactory and visual sensors. Sadly, they made another error and left the taste buds of the robot wired to the pain module. You are both given another cookie and get asked about your current level of enjoyment. Your answers are again very different, but not because the function of your enjoyment modules differs but because taste has no impact on enjoyment for the robot leaving them with a rather painful, unenjoyable experience.

Translating this thought experiment over into statistical terms is relatively straightforward. In the example, two factors (representing our latent variables of enjoyment and pain) are measured using a set of indicators. In a classical framework, one assumes that each item is uniquely related to each measured construct. In the simplest case of structural nonequivalence, individual items are related to different constructs across groups. *Structural equivalence* is the first level of equivalence that is commonly tested for with statistical methods such as Procrustes rotations and baseline fit in multigroup confirmatory factor analyses (Fontaine, 2005; Tucker, 1951).

Metric Equivalence

To continue our thought example, finally the technician has wired everything up correctly. Unfortunately, the factory producing the taste sensor has done shoddy work and it sends back substantially less signal than a comparable human taste bud. In a variation of the experiment, you are both given three cookies: one stale, one fresh, and one home-made. You can clearly taste the difference between these, and you accordingly report greater enjoyment for the fresh compared to the stale and for the home-made compared to the fresh cookie. In contrast, your robot colleague reports the same level of enjoyment for all the cookies.

Translating this again into statistical terms, in this case, while the structure underpinning our measure is equivalent and the same items load onto the same factors, they do not do so in equal strength. Violation of *metric equivalence* results in several issues for researchers, such as incomparability of relationships between constructs across groups. While some researchers have argued that the presence of meaningful nomological network across groups should be taken as an indication that valid conclusions can be made even from nonequivalent data (Welzel et al., 2021), subsequent research indicates that even a meaningful nomological network can mask statistical issues (Fischer et al., 2022).

Scalar Equivalence

Finally, all errors have been fixed in our robot, but our experiment has been going on for so long that it is now low on power, thus sending generally less intensive signals to conserve power. You are again given the same cookies. You now rate them in the same sequence as enjoyable, but your robot colleague rates all cookies as less enjoyable than you. This would mean that the ordering of cookies would be the same among yourself and the robot, but the absolute level of enjoyment would not be comparable.

This is most commonly the last level that researchers test for as fulfilment of scalar equivalence allows for direct mean comparisons between groups. The issue is that it is also the most difficult level of equivalence to achieve with natural data, with in some cases only 25% of demographic comparisons and 28% of between-group experimental designs reaching this level (D'Urso et al., 2022). One of the reasons for this is that all previous levels have to be satisfied but also because this property is vulnerable to a wide range of disturbance factors such as cultural knowledge, familiarity with the test, or general response biases. This has led researchers to either implicitly or explicitly omit tests for scalar invariance (D'Urso et al., 2022) as they might perceive it as a roadblock to their intended research (Karl & Fischer, 2022).

Approaches to Test Equivalence

Briefly considered, approaches to testing equivalence can be considered either as *exact* or *approximate*. Exact equivalence rests on the assumption that the measurement should be exactly identical across groups and any deviation should be considered an issue. In contrast, approximate equivalence rests on the assumption that the measurement between groups only needs to be approximately identical as small deviations might not meaningfully impact comparison. Both approaches have their distinct benefits and drawbacks that need careful theoretical consideration by researchers and sufficient knowledge about the appropriate tools. The following section outlines commonly used tools for each approach and highlights their benefits and issues.

Exact Equivalence Using Confirmatory Approaches

The general advantage of approaches based on confirmatory factor analysis (CFA) is that they explicitly perform an empirical test of the theoretical structure a researcher has envisioned for the target instrument. Underpinning CFAs is a theory-driven approach that aims at modeling the covariance between items and variances of individual items. In this sense, it captures a measurement model that treats items (in psychology most commonly responses to Likert-type questions) as indicators of a theoretically assumed underlying latent construct (Bollen, 1989; Long, 1983). This requires researchers to have a strong a priori expectation about which items are supposed to load on which latent variable and what the higher-order structure of the latent variables should be. While researchers have a range of different modeling

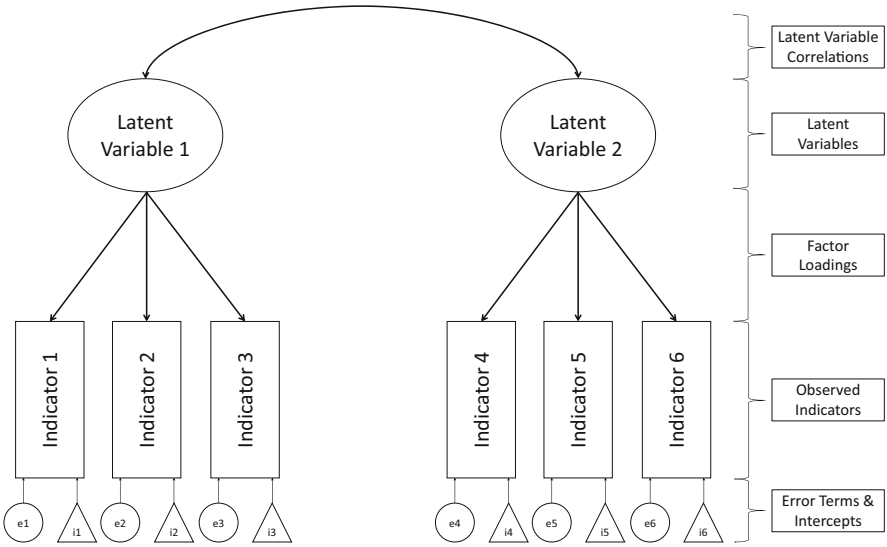

Fig. 1 Schematic representation of confirmatory factor analysis models

options for higher-order latent factors, they are probably most commonly interested in structures with one level of latent variables under which the observed indicators are subsumed (see Fig. 1 for a schematic representation). This is not to say that researchers interested in cross-group comparisons using CFA are limited to investigating these simple structures. Past research has used these methods to investigate the presence of second-order factors, general factors, and methods factors (for examples in the mindfulness literature, see Aguado et al., 2015; Karl et al., 2020; Van Dam et al., 2012). If researchers are interested in more complex arrangements of latent and observed variables, the CFA approach can be combined with exploratory structural equation modeling that allows for structures in which items show cross-loadings between constructs (Marsh et al., 2009; van Zyl & ten Klooster, 2022).

An elementary requirement of a CFA is the presence of at least three items per latent factor (for caveats, see Bollen, 1989). Additionally, one factor loading has to be set to 1 to allow for identification of the model and allow for an appropriate scaling of the latent variable. This can result in problems for cross-group comparisons as this might result in researchers choosing an item to be fixed to 1 in both groups, which is an especially poor indicator resulting in low convergence between groups. To avoid this, researchers who are not interested in the latent variable mean and variance can fix the mean of the latent variable to 0 and the variance to 1, which allows for the restraints on the factor loading to be eased while still allowing for model identification (Lance & Vandenberg, 2002). Researchers also need to make careful decisions based on their theoretical and empirical knowledge about the data, taking care to select appropriate estimation procedures. In the case of the most

commonly employed maximum likelihood approach, data need to be interval level that are also multivariate normally distributed. If researchers assume that their data do not meet the interval-level data requirement or if they observe that their data substantially diverge from the assumption of multivariate normality, a range of different estimation techniques in different statistical software packages have been developed to address these issues (Benson & Fleishman, 1994; Flora & Curran, 2004; Satorra & Bentler, 1988). It is essential that researchers exercise care when deciding their estimation approach as it can substantially impact the comparability of CFA structures and the interpretation of models between groups. Core to this is the impact of data quality and estimation on model fit (Gao et al., 2020).

As the name implies, CFA is confirmatory in nature and therefore relies heavily on established statistics of fit and misfit to determine the suitability of a hypothesized structure to the observed data. This is not to say that fit only matters for confirmatory approaches, and recent research highlights the applicability of fit measures in exploratory research (Finch, 2020). Nevertheless, as for any given data, there is a plethora of alternative models that could have been proposed by researchers, and it is essential to examine fit to the data to achieve a semblance of insight into the appropriateness of a solution.

Roughly speaking, fit indices can be divided into three different categories: *absolute*, *incremental*, and *parsimonious,* of which the absolute and incremental have probably played the biggest role in applied CFA research (Kline, 2015). Absolute fit indices examine the total misfit of the hypothesized variance–covariance matrix relative to the observed variance–covariance matrix. One common indicator in this camp is χ^2, representing the absolute mismatch between the data and the specified model that can be evaluated using common significance testing to support decisions about the acceptability of a model (Barrett, 2007). If χ^2 indicates no significant deviation from the exact fit, this would indicate that any misspecifications in the model are negligible. While this approach seems intuitive at first, it comes with a range of pitfalls. First, researchers know that each model they fit is an abstraction of reality and therefore by necessity shows nonexact fit to the data. If this would not be the case, research would fail at one of its core missions, to reduce the complexity of the observed world into simpler more general mechanisms and laws (Browne & Cudeck, 1992). Second, as all other significance-based tests against exact 0, the likelihood of rejection increases systematically with sample size as a result of minor differences becoming magnified due to increasingly tight confidence intervals (Bentler & Bonett, 1980; Bollen, 1989). This limits the applicability of the χ^2 as tool for decision-making in CFA research. To address some of these shortcomings, by, for example, considering model complexity, researchers have developed a range of additional absolute fit indices such as the goodness-of-fit index (GFI), adjusted GFI, root mean square error of approximation (RMSEA), and root mean square residual and standardized root mean square residual (SRMR). For each of these indicators, researchers have developed cut-off criteria aiming to support judgments about the acceptance or rejection of a hypothesized model. Similar to the χ^2, researchers have adopted conventional cut-off values that are thought to signal good model fit for these statistics. RMSEA is historically considered good if ranging

between 0.06 and 0.08, but simulations by Hu and Bentler (1998, 1999) suggested that a cut-off of 0.06 might be more appropriate. Similarly, for SRMR, a commonly recommended cut-off has been 0.08 (Hu & Bentler, 1999).

In contrast, incremental fit indices compare the model hypothesized by the researcher to a baseline model that specifies no relationships between the observed and latent variables and only contains variances for observed variables. This baseline model represents what could be considered an absolute null hypothesis, that the variables are not meaningfully related. Incremental fit indices represent the improved fit for the model compared to this absolute null model. Examples are comparative fit index (CFI), normed-fit index (NFI), and non-normed fit index (Bentler, 1990; Bentler & Bonett, 1980). Higher values indicate better fit, where values above 0.95 can be interpreted as a good fit (Hu & Bentler, 1998).

Finally, parsimonious fit indices represent an expansion of the previous fit indices by adjusting the observed fit for the number of parameters added. Similar to other covariance-based indicators (such as α), CFA fit indices tend to show better fit with increasing number of parameters such as paths or loadings. Parsimonious fit indices address this trend toward ideal fit by penalizing the fit with increasing model complexity. Examples are the parsimony goodness-of-fit index (Mulaik et al., 1989) and the parsimony normed fit index (James et al., 1982). Importantly, while each of the different fit indicators yield a different perspective on model fit, none of them should be used exclusively as this might lead researchers to accept a model that might fit well from one perspective but might show poor fit based on other indicators. Further, model fit in itself is not a goal, and researchers, if they modify their theoretically specified model for example by adding item covariances or cross-loadings, should be mindful to not neglect theory in pursuit of better model fit, which can be difficult to achieve across the board as each indicator aims to capture different aspects of misspecification.

Multigroup confirmatory factor analysis (MGCFA) is probably the most widely used approach to test measurement invariance between groups. MGCFA is a statistical technique that builds upon the foundations of confirmatory factor analysis. In multigroup confirmatory factor analysis, the parameters of the model are constrained to be equal across groups, providing a straightforward way to examine how differences between groups at different levels of the model influence model fit. This stepwise approach is often used to test for structural, metric, and scalar equivalence, with each level representing a separate step. Researchers may also test for the equivalence of means and standard errors, although this is less common as most researchers are primarily interested in comparing correlations or means across countries, for which metric and scalar equivalence are typically sufficient. By conducting MGCFA, researchers can gain insight into how cultural, linguistic, or other differences between groups may impact their responses to survey items or measures, which can be important for understanding and comparing the validity of research findings across different populations. Commonly, researchers first examine the fit of a model across cultures in which all parameters except the patterns of loadings are allowed to vary freely across cultures. Sufficient fit according to the commonly used fit criteria outlined before is taken as evidence that across all

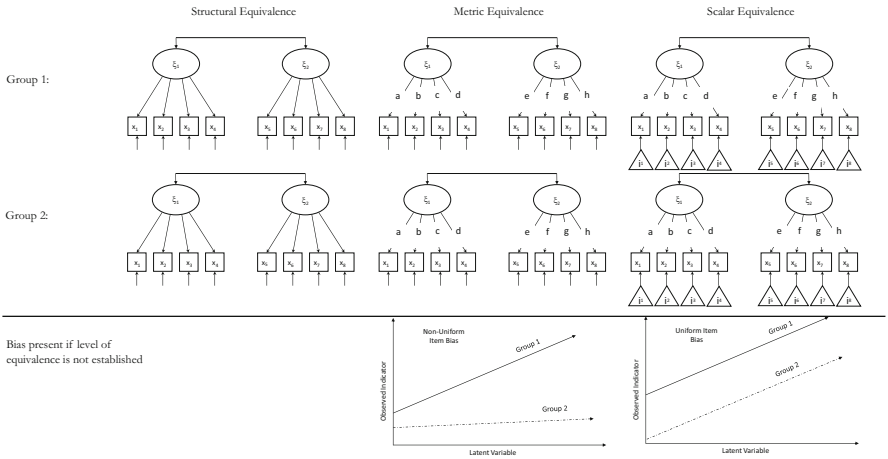

Fig. 2 Levels of measurement equivalence and the impact on comparability between groups

cultures the structure of the model is not substantially miss-specified providing evidence for structural equivalence. Next, researchers restrict the factor loadings on each factor to be equal across groups and examine the drop in fit from the model in which only the structure was constrained to this loading constrained model. If the drop is below a certain criterion, for example, <0.01 for CFI, this is taken as evidence that the countries share an overall similar pattern of loadings (Kang et al., 2016; Little, 1997; for an empirical argument to relax this threshold, see Rutkowski & Svetina, 2014), supporting metric equivalence. Last, researchers constrain the item intercepts to be identical across groups and compare this intercept-restricted model to the loadings-restricted model, again a drop below a certain predefined cut-off such as <0.01 for CFI is taken as indication of identical intercepts, supporting scalar equivalence (Fig. 2 shows a representation of these different models and their implications for cross-group bias).

While testing for multigroup invariance is a valuable technique that is approachable from a wide range of software solutions, it is important to recognize that this procedure comes with its own difficulties and intricacies. One of the primary challenges is interpreting noninvariance. Some researchers view noninvariance as an indication that meaningful research cannot be conducted, and tests of invariance should be disregarded (for a commentary on this, see Fischer et al., 2022; Welzel et al., 2021). However, a more productive approach is to identify the sources of misfit and explore the reasons behind the lack of invariance. One potential issue that researchers face when testing for multigroup invariance is the complexity of the model. Parceling, which involves randomly assigning items to mean parcels, has been used to address this issue. However, while this approach may improve the fit of the model, it often does not address the underlying issues of noninvariance and can instead hide them. Therefore, this technique should be discouraged as it does not only come with its own statistical problems, but, because these mean parcels represent latent variables in themselves, it essentially devalues the approach.

Assuming that structural equivalence has been established, the next common stumbling block is metric and scalar equivalence. In the past, nonequivalence at these levels has been addressed using modification indices to identify constrained parameters between groups that could be released to reduce the misfit between the observed data and the specified model. While this approach can be helpful in identifying parts of the model that may not be invariant between groups and selectively allowing them to vary, it raises theoretical questions regarding how many parameters can be released before no equivalence test is being conducted anymore (Byrne et al., 1989; Meredith, 1993; Shi et al., 2019). Additionally, it is unclear how the freed indicators should be treated in subsequent model, especially if the subsequent model uses observed rather than latent scores, or what their impact on group comparisons is.

One way that has recently been developed to address these issues is differential multiple-group analysis of covariance structures (DMACS). This is a statistical technique used in multigroup invariance testing to compare the factor structure of a measurement instrument across different groups. In addition to testing for differences in factor structure, DMACS can also provide information on effect sizes (Gunn et al., 2020), which can help researchers interpret the practical significance of the differences observed between group intercepts and loadings. Effect size measures provide information about the magnitude of the differences between groups, which can help researchers determine whether the differences observed are practically significant. In DMACS, effect sizes can be calculated for each comparison of factor structures between groups using measures such as Cohen's *d* or the Omega squared coefficient. While this yield informative indicators for within-model comparisons, interpreting DMACS effect sizes across models can be challenging as there is no clear consensus on what constitutes a "large" or "small" effect size in this context. However, some guidelines suggest that effect sizes of around 0.20 may be considered small, while effect sizes of around 0.50 or higher may be considered large. Effect sizes can also be used to compare the magnitude of differences across different factors or dimensions being measured (Karl & Fischer, 2022). For example, if one factor has a large effect size average difference across items while another has a small effect size, this can help researchers prioritize which factors are most important to focus on in further research. DMACS effect sizes can present a useful tool for interpreting the magnitude of these differences and prioritizing areas for further investigation. By providing a more nuanced understanding of the differences between groups, effect sizes can help researchers develop more accurate and meaningful conclusions about the invariance of measurement instruments across multiple groups.

Finally, it is important to consider the sample size on the group level when testing for multigroup invariance. Although a large number of groups (e.g., more than 20, Asparouhov & Muthén, 2014) can provide more reliable results, it can also increase the likelihood of finding small differences that may not be practically significant. Therefore, it is important to carefully consider the sample size when conducting invariance testing. In conclusion, while testing for multigroup invariance using alignment is a valuable technique, it is important to be aware of the potential difficulties and intricacies involved. Researchers should take care to interpret non-invariance and identify sources of misfit, address potential issues such as item bias,

and carefully consider the sample size when conducting invariance testing. By doing so, researchers can ensure that their measurement instruments are valid and reliable across different groups, allowing for meaningful comparisons and insights.

Approximate Equivalence

Up until now this chapter has been focusing on an approach that has been labeled *exact equivalence*. In this view, equivalence exists if there is no deviation between parameters of interest across groups. Contrasting this view is *approximate equivalence* according to which parameters only need to show near identity across groups and small differences should be discarded. This approach has been used to extend MGCFA to multigroup factor analysis alignment (from here on in this chapter referred to as *alignment*; Asparouhov & Muthén, 2014). Alignment is a relatively new approach to testing group differences, which has gained popularity in recent years (Muthén & Asparouhov, 2018). It is based on the idea that different groups may have different ways of conceptualizing and measuring the same construct, but these differences may still be compatible with each other. For example, two groups may have different items or indicators for measuring a construct, but these items may still be measuring the same underlying concept.

Alignment proceeds in two steps, which can be described as fitting and optimization of a model (for an illustrative example, see Luong & Flake, 2022). In the first step, latent factor means are fixed to 0 and latent factor variances are fixed to 1, resulting in an initial configural model. In the second step, this model is optimized using a component loss function that aims to minimize the noninvariance in means and factor variances for each group (for a detailed mathematical description, see Asparouhov and Muthén, 2014). This optimization process terminates at a point at which "there are few large non-invariant measurement parameters and many approximately non-invariant parameters rather than many medium-sized non-invariant measurement parameters" (Asparouhov & Muthén, 2014, p. 497). Overall, this approach has been shown to yield robust results, with recent work demonstrating improved algorithms (Pokropek et al., 2020; Robitzsch, 2020).

Overall, in alignment, researchers aim to identify the commonalities between groups and find a way to reconcile the differences that can allow for comparable means even in the presence of noninvariance. The basic idea behind alignment is that different groups may have different factor structures, but these structures may be aligned with each other in a meaningful way. Alignment involves finding a set of "anchor items" that are common across groups and using these items to establish a common metric for the construct (this could be seen as similar to item banking or a less radical version of recent calls to relax the need for semantic similarity of items; Boehnke, 2022). The anchor items are chosen based on their high factor loadings and high correlations with the construct. Once the anchor items are identified, researchers can use them to establish a common metric for the construct and compare the scores of different groups on this metric.

Alignment has several advantages over MGCFA. One advantage is that it does not require exact measurement equivalence between groups. In MGCFA, researchers assume that the factor structure and measurement parameters are exactly the same across groups, which is often unrealistic especially if a large number of groups is considered. In contrast, alignment focuses on identifying the commonalities. It allows for some differences between groups, with around 20% noninvariance representing the threshold for problematic noninvariance in cross-sectional and longitudinal contexts (Asparouhov & Muthén, 2014; Lai, 2023). This makes alignment more flexible and robust to violations of the measurement equivalence assumption. Alignment also has some limitations that researchers should be aware of. One limitation is that it may be difficult to identify the anchor items, especially if the groups have very different factor structures or measurement parameters. In some cases, researchers may need to collect additional data to identify the anchor items or use expert judgment to select the items. Additionally, the method is still relatively niche among researchers, but software solutions are available both in Mplus and R, which might allow for a wider implementation.

Conclusion

Research in behavioral health is becoming increasing global, recognizing the limited perspective in the past literature where the majority of research focused on WEIRD populations (Western, Educated, Industrialized, Rich, and Democratic; Henrich, 2020). With this increase in global and cross-cultural research comes the necessity to establish solid measurement foundations that allow researchers to make firm conclusions about the populations that they want to compare. This chapter highlighted the underpinning theoretical foundations behind measurement equivalence and highlighted selected methods, which allow researchers to test for measurement equivalence in their data. This is not to say that these are the only methods. As can be seen with the development of alignment, the field is in consistent development. Researchers are encouraged to stay mindful of these developments if they are interested in performing cross-group work and eschew ritualistic testing of noninvariance in favor of critically considering what noninvariance implies for their field of study and their data.

References

Aguado, J., Luciano, J. V., Cebolla, A., Serrano-Blanco, A., Soler, J., & García-Campayo, J. (2015). Bifactor analysis and construct validity of the five facet mindfulness questionnaire (FFMQ) in non-clinical Spanish samples. *Frontiers in Psychology, 6*, 404. https://doi.org/10.3389/fpsyg.2015.00404

Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling, 21*(4), 495–508. https://doi.org/10.1080/10705511.2014.919210

Atari, M., Haidt, J., Graham, J., Koleva, S., Stevens, S. T., & Dehghani, M. (2023). Morality beyond the WEIRD: How the nomological network of morality varies across cultures. *Journal of Personality and Social Psychology.* https://doi.org/10.1037/pspp0000470

Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual Differences, 42*(5), 815–824. https://doi.org/10.1016/j.paid.2006.09.018

Benson, J., & Fleishman, J. A. (1994). The robustness of maximum likelihood and distribution-free estimators to non-normality in confirmatory factor analysis. *Quality and Quantity, 28*(2), 117–136. https://doi.org/10.1007/BF01102757

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*(3), 588–606. https://doi.org/10.1037/0033-2909.88.3.588

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238–246. https://doi.org/10.1037/0033-2909.107.2.238

Boehnke, K. (2022). Let's compare apples and oranges! A plea to demystify measurement equivalence. *American Psychologist, 77*(9), 1160–1168. https://doi.org/10.1037/amp0001080

Boer, D., Hanke, K., & He, J. (2018). On detecting systematic measurement error in cross-cultural research: A review and critical reflection on equivalence and invariance tests. *Journal of Cross-Cultural Psychology, 49*, 713–734. https://doi.org/10.1177/0022022117749042

Bollen, K. A. (1989). *Structural equations with latent variables*. John Wiley & Sons. https://doi.org/10.1002/9781118619179

Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology, 53*, 605–634. https://doi.org/10.1146/annurev.psych.53.100901.135239

Borsboom, D. (2008). Latent variable theory. *Measurement: Interdisciplinary Research and Perspectives, 6*, 25–53. https://doi.org/10.1080/15366360802035497

Borsboom, D., Deserno, M. K., Rhemtulla, M., Epskamp, S., Fried, E. I., McNally, R. J., Robinaugh, D. J., Perugini, M., Dalege, J., Costantini, G., Isvoranu, A.-M., Wysocki, A. C., van Borkulo, C. D., van Bork, R., & Waldorp, L. J. (2021). Network analysis of multivariate data in psychological science. *Nature Reviews Methods Primers, 1*, 1–18. https://doi.org/10.1038/s43586-021-00055-w

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research, 21*(2), 230–258. https://doi.org/10.1177/0049124192021002005

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*(3), 456–466. https://doi.org/10.1037/0033-2909.105.3.456

Christopher, M. S., Charoensuk, S., Gilbert, B. D., Neary, T. J., & Pearce, K. L. (2009). Mindfulness in Thailand and the United States: A case of apples versus oranges? *Journal of Clinical Psychology, 65*, 590–612. https://doi.org/10.1002/jclp.20580

D'Urso, E. D., Maassen, E., van Assen, M. A. L. M., Nuijten, M. B., Roover, K. D., & Wicherts, J. (2022). The dire disregard of measurement invariance testing in psychological science. *PsyArXiv.* https://doi.org/10.31234/osf.io/n3f5u

Finch, W. H. (2020). Using fit statistic differences to determine the optimal number of factors to retain in an exploratory factor analysis. *Educational and Psychological Measurement, 80*(2), 217–241. https://doi.org/10.1177/0013164419865769

Fischer, R., Karl, J. A., Fontaine, J. R. J., & Poortinga, Y. H. (2022). Evidence of validity does not rule out systematic bias: A commentary on nomological noise and cross-cultural invariance. *Sociological Methods & Research.* https://doi.org/10.1177/00491241221091756

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9*(4), 466–491. https://doi.org/10.1037/1082-989X.9.4.466

Fontaine, J. (2005). Equivalence. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement* (Vol. 1, pp. 803–813).

Gao, C., Shi, D., & Maydeu-Olivares, A. (2020). Estimating the maximum likelihood root mean square error of approximation (RMSEA) with non-normal data: A Monte-Carlo study. *Structural Equation Modeling, 27*(2), 192–201. https://doi.org/10.1080/10705511.2019.1637741

Gunn, H. J., Grimm, K. J., & Edwards, M. C. (2020). Evaluation of six effect size measures of measurement non-invariance for continuous outcomes. *Structural Equation Modeling, 27*(4), 503–514. https://doi.org/10.1080/10705511.2019.1689507

Henrich, J. (2020). *The WEIRDest people in the world: How the west became psychologically peculiar and particularly prosperous*. Farrar, Straus and Giroux.

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3*(4), 424–453. https://doi.org/10.1037/1082-989X.3.4.424

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hui, C. H., & Triandis, H. C. (1985). Measurement in cross-cultural psychology: A review and comparison of strategies. *Journal of Cross-Cultural Psychology, 16*(2), 131–152. https://doi.org/10.1177/0022002185016002001

James, L. R., Mulaik, S. A., & Brett, J. M. (1982). *Causal analysis: Assumptions, models, and data* (1st ed.). Sage Publications.

Jeong, S., & Lee, Y. (2019). Consequences of not conducting measurement invariance tests in cross-cultural studies: A review of current research practices and recommendations. *Advances in Developing Human Resources, 21*, 466–483. https://doi.org/10.1177/1523422319870726

Kang, Y., McNeish, D. M., & Hancock, G. R. (2016). The role of measurement quality on practical guidelines for assessing measurement and structural invariance. *Educational and Psychological Measurement, 76*(4), 533–561. https://doi.org/10.1177/0013164415603764

Karl, J. A., & Fischer, R. (2022). More than yes and no: Predicting the magnitude of non-invariance between countries from systematic features. In *Proceedings of the IACCP20+*. IACCP 20+.

Karl, J. A., Méndez Prado, S. M., Gračanin, A., Verhaeghen, P., Ramos, A., Mandal, S. P., Michalak, J., Zhang, C.-Q., Schmidt, C., Tran, U. S., Druica, E., Solem, S., Astani, A., Liu, X., Luciano, J. V., Tkalčić, M., Lilja, J. L., Dundas, I., Wong, S. Y. S. Y., & Fischer, R. (2020). The cross-cultural validity of the five-facet mindfulness questionnaire across 16 countries. *Mindfulness, 11*(5), 1226–1237. https://doi.org/10.1007/s12671-020-01333-6

Kline, R. B. (2015). *Principles and practice of structural equation modeling* (4th ed.). Guilford Press.

Kroenke, K., & Spitzer, R. L. (2002). The PHQ-9: A new depression diagnostic and severity measure. *Psychiatric Annals, 32*, 509–515. https://doi.org/10.3928/0048-5713-20020901-06

Kroenke, K., Spitzer, R. L., Williams, J. B. W., Monahan, P. O., & Löwe, B. (2007). Anxiety disorders in primary care: Prevalence, impairment, comorbidity, and detection. *Annals of Internal Medicine, 146*, 317–325. https://doi.org/10.7326/0003-4819-146-5-200703060-00004

Lai, M. H. C. (2023). Adjusting for measurement noninvariance with alignment in growth modeling. *Multivariate Behavioral Research, 58*(1), 30–47. https://doi.org/10.1080/00273171.2021.1941730

Lance, C. E., & Vandenberg, R. J. (2002). Confirmatory factor analysis. In F. Drasgow & N. Schmitt (Eds.), *Measuring and analyzing behavior in organizations: Advances in measurement and data analysis* (pp. 221–254). Jossey-Bass/Wiley.

Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research, 32*(1), 53–76. https://doi.org/10.1207/s15327906mbr3201_3

Long, J. S. (1983). *Covariance structure models: An introduction to LISREL*. SAGE Publications.

Luong, R., & Flake, J. K. (2022). Measurement invariance testing using confirmatory factor analysis and alignment optimization: A tutorial for transparent analysis planning and reporting. *Psychological Methods.* https://doi.org/10.1037/met0000441

Marsh, H. W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J. S., & Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Application to

students' evaluations of university teaching. *Structural Equation Modeling, 16*(3), 439–476. https://doi.org/10.1080/10705510903008220

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525–543. https://doi.org/10.1007/BF02294825

Mulaik, S. A., James, L. R., Van Alstine, J., Bennett, N., Lind, S., & Stilwell, C. D. (1989). Evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin, 105*(3), 430–445. https://doi.org/10.1037/0033-2909.105.3.430

Muthén, B., & Asparouhov, T. (2018). Recent methods for the study of measurement invariance with many groups: Alignment and random effects. *Sociological Methods & Research, 47*(4), 637–664. https://doi.org/10.1177/0049124117701488

Pokropek, A., Lüdtke, O., & Robitzsch, A. (2020). An extension of the invariance alignment method for scale linking. *Psychological Test and Assessment Modeling, 62*, 305–334.

Robitzsch, A. (2020). Lp loss functions in invariance alignment and Haberman linking with few or many groups. *Stat, 3*(3), 246–283. https://doi.org/10.3390/stats3030019

Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement, 74*(1), 31–57. https://doi.org/10.1177/0013164413498257

Satorra, A., & Bentler, P. M. (1988). Scaling corrections for chi-square statistics in covariance structure analysis. In *ASA 1988 proceedings of the business and economic statistics section* (pp. 308–313).

Shi, D., Song, H., & Lewis, M. D. (2019). The impact of partial factorial invariance on cross-group comparisons. *Assessment, 26*(7), 1217–1233. https://doi.org/10.1177/1073191117711020

Simons, R. C., & Hughes, C. C. (Eds.). (1985). *The culture-bound syndromes: Folk illnesses of psychiatric and anthropological interest: 7*. Springer.

Tseng, W.-S. (2006). From peculiar psychiatric disorders through culture-bound syndromes to culture-related specific syndromes. *Transcultural Psychiatry, 43*(4), 554–576. https://doi.org/10.1177/1363461506070781

Tucker, L. R. (1951). A method for synthesis of factor analysis studies. Personnel Research Report Department Army., No 984.

Van Dam, N. T., Hobkirk, A. L., Danoff-Burg, S., & Earleywine, M. (2012). Mind your words: Positive and negative items create method effects on the five facet mindfulness questionnaire. *Assessment, 19*(2), 198–204. https://doi.org/10.1177/1073191112438743

van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Sage Publications.

Van Doren, N., Zainal, N. H., & Newman, M. G. (2021). Cross-cultural and gender invariance of emotion regulation in the United States and India. *Journal of Affective Disorders, 295*, 1360–1370. https://doi.org/10.1016/j.jad.2021.04.089

van Zyl, L. E., & ten Klooster, P. M. (2022). Exploratory structural equation modeling: Practical guidelines and tutorial with a convenient online tool for Mplus. *Frontiers in Psychiatry, 12*, 795672. https://www.frontiersin.org/articles/10.3389/fpsyt.2021.795672

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4–70. https://doi.org/10.1177/109442810031002

Waldren, L. H., Livingston, L. A., & Shah, P. (2022). Is there an optimal self-report measure to investigate autism-related sex differences? *PsyArXiv*. https://doi.org/10.31234/osf.io/r4t9v

Wang, M., Summers, J. A., Little, T., Turnbull, A., Poston, D., & Mannan, H. (2006). Perspectives of fathers and mothers of children in early intervention programmes in assessing family quality of life. *Journal of Intellectual Disability Research, 50*, 977–988. https://doi.org/10.1111/j.1365-2788.2006.00932.x

Welzel, C., Brunkert, L., Kruse, S., & Inglehart, R. F. (2021). Non-invariance? An overstated problem with misconceived causes. *Sociological Methods & Research*. https://doi.org/10.1177/0049124121995521

Wu, P.-C., & Huang, T.-W. (2014). Gender-related invariance of the Beck depression inventory II for Taiwanese adolescent samples. *Assessment, 21*(2), 218–226. https://doi.org/10.1177/1073191112441243

Zager Kocjan, G., Jose, P. E., Sočan, G., & Avsec, A. (2021). Measurement invariance of the subjective happiness scale across countries, gender, age, and time. *Assessment, 29*(4), 826–841. https://doi.org/10.1177/1073191121993558