Article

Measurement Invariance Testing Works

Applied Psychological Measurement
2024, Vol. 48(6) 257–275

DOI: 10.1177/01466216241261708
journals.sagepub.com/home/apm
Sage

Jordan Lasker[1]

Abstract

Psychometricians have argued that measurement invariance (MI) testing is needed to know if the same psychological constructs are measured in different groups. Data from five experiments allowed that position to be tested. In the first, participants answered questionnaires on belief in free will and either the meaning of life or the meaning of a nonsense concept called "gavagai." Since the meaning of life and the meaning of gavagai conceptually differ, MI should have been violated when groups were treated like their measurements were identical. MI was severely violated, indicating the questionnaires were interpreted differently. In the second and third experiments, participants were randomized to watch treatment videos explaining figural matrices rules or task-irrelevant control videos. Participants then took intelligence and figural matrices tests. The intervention worked and the experimental group had an additional influence on figural matrix performance in the form of knowing matrix rules, so their performance on the matrices tests violated MI and was anomalously high for their intelligence levels. In both experiments, MI was severely violated. In the fourth and fifth experiments, individuals were exposed to growth mindset interventions that a twin study revealed changed the amount of genetic variance in the target mindset measure without affecting other variables. When comparing treatment and control groups, MI was attainable before but not after treatment. Moreover, the control group showed longitudinal invariance, but the same was untrue for the treatment group. MI testing is likely able to show if the same things are measured in different groups.

Keywords

measurement invariance, theory-testing, psychometrics, group differences, comparisons, modeling, SEM

[1]Texas Tech University, Lubbock, TX, USA

Corresponding Author:
Jordan Lasker, Texas Tech University, 2500 Broadway W, Lubbock, TX 79409, USA.
Email: jlasker@ttu.edu

Introduction

> [I]f the error variances are different then there are either different variables operating on the measure across groups or the same set of variables operate differently across groups. There simply is no alternative explanation. (DeShon, 2004, p. 144)

The principal aim of measurement invariance (MI) testing is assessing whether a given psychological instrument measures the same thing in different groups. This is typically assessed through a process in which models are iteratively fitted until factor configurations, loadings, and intercepts, are equated between groups.[1] The interpretation of equal factor configurations (configural invariance) is that groups reference the same structures of constructs when dealing with the items on a given psychological measure. Equal loadings (metric invariance) are interpreted to mean that groups think of psychological measures the same way when they provide their answers. Equal intercepts (scalar invariance) imply the levels of psychological measures can be considered to represent the levels of a measurement's target construct to the same degree.

These three steps have been defended as all that is needed to achieve MI (Little et al., 2007). A major component that is frequently disregarded is the residual variances—the equating of which is known as "strict (factorial) invariance." Strict invariance is generally agreed to be the last stage of MI testing and subsequent constraints fall under what is known as structural invariance testing. Though these concepts are related, they are not the same (Vandenberg & Lance, 2000). If the conditions of MI testing are fulfilled up to scalar invariance, we can compare the means of latent variables and make the statement that performance differences between groups are driven systematically by differences in the levels of those latent variables. If we achieve strict invariance, we can say that a psychological instrument measures the same things between populations because all influences, including ones that are not explicitly modeled, are equated during that step.[2]

There are many publications that describe how to test MI (Fischer & Karl, 2019; Putnick & Bornstein, 2016; van de Schoot et al., 2012). Lasker (2021), for example, illustrated each step of conducting a multi-group confirmatory factor analysis (MGCFA) and computing the impact of MI violations with attached code. There are many methods besides MGCFA for detecting violations of MI (Millsap, 2011; Penfield & Camilli, 2006), but MGCFA is the method this article focuses on. The ordered stages to assess MI with it are shown in Table 1, with the additional constraints added to in each stage listed. Ascertaining whether the different stages have been met requires contrasting the model's fit in each subsequent stage with the fit from the prior stage.

Table 1. Steps for Multi-Group Confirmatory Factor Analytic Measurement Invariance Testing.

Stage	Constraints	Comparisons allowed
1 Configural invariance	The same model fits in both groups	None
2 Metric invariance	1 + loadings are constrained to equality between groups. Latent variances and covariances can still differ between groups	Latent variances and covariances (potentially)
3 Scalar invariance	2 + intercepts are constrained to equality between groups. Latent means are allowed to vary between groups	Latent means, variances, and covariances (very likely)
4 Strict invariance	3 + error variances are constrained to equality between groups	

Note. This table is adapted from Beaujean (Beaujean, 2014, tbl. 4.1). For additional information, see Beaujean (2014, chp. 4).

It is important to recognize that MI, taken as the fulfillment of the three steps prior to testing strict invariance[3] does *not* allow us to state that we are measuring the same things in different groups. Scalar invariance alone is still consistent with a scenario in which unmeasured variables drive differences in scoring (Lubke et al., 2003), and it can mask differences in specific indicator means, affecting our ability to trust our manifest and—consequently—latent scaling (Lubke & Dolan, 2003). Additionally, MI short of strict invariance is consistent with a scenario in which tests (and factors) are unequally reliable in different groups. Early practitioners of MI testing and its methodological ancestors were aware of this fact and what it implied for measurement (Cronbach, 1947; Meredith, 1993); more recently, this fact has also been explicitly articulated:

> "Following Cronbach's (1947) statement on error, if the error variances are different then there are either different variables operating on [a] measure across groups or the same set of variables operate differently across groups. There simply is no alternative explanation…. Given this knowledge, it is completely unclear why it would ever be acceptable to conclude that measures are invariant when the error variances are not homogeneous across groups….
>
> "Error variance is not only a random process, but also the effect of unmodeled sources of systematic variance that affect measured responses. The effect of those unmodeled sources of systematic variance on the evaluation of MI was demonstrated for the single indicator case. In those data, the measurement process is systematically different across groups and yet this difference is only manifested in the existence of error variance heterogeneity. The expected value for the response across groups (e.g., intercept and slope) was invariant…. [M]ulti-indicator models do not reduce the need to demonstrate error variance homogeneity when assessing the functioning of a measure across groups." (DeShon, 2004, pp.'s 144, 147–148).

While it is incorrect to say that invariance of the loadings and intercepts of a measurement model indicates that the same thing(s) is (are) necessarily being measured in different groups, their invariance does indicate that some of the same thing(s) is (are) being measured. Such a finding does not make strong guarantees, however. On its own, the mere achievement of this level of MI ignores potential differences in causes specific to one or another group. To the extent reliability differs, so too must the causes of responses and what is measured in different groups, regardless of the random or systematic nature of the residual variances. Consider Cronbach's (1947) statement that "All methods of studying reliability make a somewhat fallacious division of variables into 'real variables' and 'error.' [….] A test score is made up of all of these 'real' elements, each of which could be perfectly predicted if our knowledge were adequate. Reliability, according to this conception, becomes a measure of our ignorance of the real factors underlying fluctuations of behavior and atypical acts." (ibid., pp. 6–7). If we had perfect knowledge, we would know exactly what causes "error"; if we know we have strict invariance, then when it comes to group comparisons, we don't need to.

MI is clearly a very important concept for testing theories and performing psychological comparisons in general. This importance has been recognized in numerous papers (Boer et al., 2018; Jeong & Lee, 2019; Lacko et al., 2022; Lasker et al., 2022; Wicherts, 2016), and yet confusion about (Welzel et al., 2021; cf. Meuleman et al., 2022) and disregard of it (Maassen et al., 2023) remain the norm. This may reflect a lack of appreciation for what MI, or its violation, can reveal to researchers. As Maassen et al. (2023) argued, MI testing should be "an opportunity for researchers to gain a clearer understanding of the substantive and practical meaning of the differences they observe" (ibid., p. 12).

Several recent papers have provided data from situations that have made it possible to empirically test whether MI represents measuring the same things in different groups (Burgoyne

et al., 2020; Protzko, 2022[4]; Schneider et al., 2020). Before directing our attention to them, several things to keep in mind when testing MI must be noted.

The interpretation of MI without strict invariance is not necessarily one of common measurement. It is, at best, one of partially common measurement and is thus not a theoretically applicable test of whether MI involves common measurement. Second, many model fit indices in common use are often insufficient to detect violations of any level of MI. Protzko (2022), for example, exclusively used two approximate fit indices, RMSEA and CFI. These are less powerful for testing omnibus MI violations than χ^2 tests or the use of information-theoretic indices like the Akaike and Schwarz' information criteria (AIC and BIC, respectively). However, they are clearly much more common (Ropovik, 2015, see, esp., the sections "Approximate Fit Indices" and "The Consequences of Disregarding the Model Test"), perhaps *because* of their weaknesses. Namely, these model fit indices may allow researchers to "null hack" (Protzko, 2018), avoiding finding evidence of misspecification or other group differences in model parameters. Additionally, many purported tests of MI feature liberal fit criteria (see Svetina et al., 2020, Table 1). This is not to say that χ^2 tests and indices like AIC and BIC are perfect, but they are known to be more sensitive means of detecting misfit.

Researchers should exercise considerable caution when using approximate fit indices because they are often insensitive, and their cutoffs are usually arbitrary.[5] This should not be taken as a complete condemnation of their use or a statement about their inherent inferiority compared to other tests. $\chi 2$ tests and information-theoretic indices have their own downside, in that they are more likely to highlight noninvariance at large sample sizes when there is *practically* none, forcing MI testing to be a matter of degree.

Another major issue in MI testing is theoretical: to understand when and why MI is expected to hold or fail, theoretical knowledge about our measurements and samples is required. Individuals from a culture valuing toughness may never supply accurate answers to questions about pain. Men may misreport their heights and weights to a greater degree and in a different direction from women because it's socially desirable for them to be tall and muscular. An English second language learner may fail to show their abilities on a test given in English due less to ability and more to lacking familiarity with the language. Each of these things can be predicted theoretically and verified by testing for MI, but the reasons may not be simple.

An example of a complex theoretical background comes from Protzko (2022). The paper's explicit aim was to test whether the same constructs were being measured in different groups. This was done by randomly assigning a population-representative sample of 1,500 American adults to answering a questionnaire on the meaning of life or an otherwise-identical one on the meaning of an undefined nonsense concept known as gavagai, and then testing MI with respect to the items of those questionnaires despite them being different. For the purposes of testing whether a "nonsense scale" like the meaning of gavagai scale could have predictive validity, the sample was also administered a questionnaire about their beliefs in free will.

Whatever the "meaning of gavagai" questionnaire measured, it might have been the same, or similar enough, to the construct underlying the "meaning of life" questionnaire. Because of their similarities (the questionnaires use almost the same items excepting the replacement of the words "gavagai" and "life"; the question wording can be found in Table S2), we should expect some comparability in measurement, while their differences ought to elicit widespread measurement noninvariance to the extent the constructs measured truly differ. Based on how much similarity there is in the influences on these tests, we can hypothesize accordingly about the identity of the residual variances. If the responses on the questionnaires were influenced by different things because people consistently interpreted the meaning of life in a way that differed from the meaning of gavagai—despite its possible consistency within individuals—then it should be the case that the residual variances are different.

Finally, the hypothesis of common measurement often has less to do with loadings and intercepts, and more to do with residual variances (i.e., strict factorial invariance). If an investigator's aim is to see if tests measure the same things, one can never forget to equate all the measurements that have to do with influences—or, in other words, the representations of "the same things." The residual or error variances *are* latent variables, and they do represent both systematic and plausibly random causes which could be explored. If we aim to test the hypothesis that MI means the same things are being measured in different groups, we must test the hypothesis suitably, by attempting to equate the residual variances. Since the concept of something like "gavagai" is different from the concept of "life," its meaning and related responses could plausibly be considered different; if we do not measure properly, we will fail to note that, and if we assume their differences are only in common latent variables rather than the potentially uncommon residual variances, we have lost sight of the objective.

The Present Study

This analysis includes three studies.

The first study is based on the assertion that Protzko (2022) constituted an incorrect test of the important claim that MI indicates that a psychological instrument measures the same thing in different groups. Protzko (2022) utilized insensitive fit indices, in effect ignoring violations of MI; the construct identity and validity of the "meaning of gavagai" questionnaire was improperly considered, and finally; a requisite test of MI—assessing strict factorial invariance—was not conducted.

To address the paper's methodological problems, I utilized the data from Protzko (2022) to suitably assess MI. The results of this reanalysis are presented didactically, to illustrate the errors in the order they should have been observed and to lay the groundwork for four other analyses of the two other studies.

The second study was based on data from Schneider et al. (2020). These researchers performed two experiments in which they taught participants the rules of figural matrices and then assessed the effects on their test scores. In Experiment 1, 112 German university students were randomized to either a treatment or a control group. Individuals in the treatment group were shown a video in which six rules that were relevant to the subsequent figural matrix tests were explained to them, and they were also given text-based and graphical, step-by-step instructions. The control group was given instructions that were similar, but the content they received information on was about diet and nutrition and was thus not test-relevant. Participants then took four subtests of the *Leistungsprüfsystem 2* (LPS-2) involving general knowledge, marking incorrect numbers in numerical sequences, quickly adding numbers, and mental rotation, respectively. They followed this up by taking two variants of the DESIGMA-Advanced figural matrices tests. In Experiment 2, 229 German university students underwent a similar procedure, with the same intelligence test minus its addition subtest, and with only a single variant of DESIGMA-Advanced. Additional information and summary statistics can be found in Schneider et al.'s (2020) study.

In Schneider et al.'s (2020) intervention, instruction was only provided for figural matrices. Since the knowledge required to perform well on matrix tests is not like the knowledge required to do well on the other LPS-2 subtests, there should only have been an effect on matrix test performance, and not on the other subtests. There should also have been bias when comparing the treatment and control groups since the control group lacked the novel influence of explicitly test-relevant information.

The third study was built around the data from Burgoyne et al. (2020). These researchers recruited twins from the Michigan State University Twin Registry study, the Twin Study of Behavioral and Emotional Development in Children, and the Michigan Twins Project registry.

They obtained zygosity and demographic information, and measured growth mindset, grit, locus of control, cognitive ability, and a handful of other measures, before randomizing pairs of participants to either a growth mindset intervention or an active control. The growth mindset intervention involved presenting participants with content suggesting that "the brain is like a muscle—it gets stronger (and smarter) when you exercise it," whereas the active control intervention involved telling participants about the human brain with less aspirational and more factually-focused content like "the parietal lobe is where the brain interprets the sense of touch" (ibid., pp. 5–6).

The effect of Burgoyne et al.'s (2020) intervention was to augment the level of participant growth mindset without seeming to affect their other measures. More importantly, these authors used their twins to fit behavior genetic ACE models to estimate the additive genetic (A), shared environmental (C), and nonshared environmental (E) variance in their mindset measures before and after their intervention, by group (Plomin et al., 2008). Their intervention increased the A variance in their measure of growth mindset in the intervention group, while at the same time, not affecting the A, C, or E variance in a self-determination composite score that they defined in their earlier pilot study (Burgoyne et al., 2018) as a score for a higher order construct called "self-determination," composed of responses for growth mindset, grit, and locus of control.

Burgoyne et al.'s (2020) finding is perfect for testing whether MI means the same things are being measured in different groups. This is because they found that responding with their mindset measure involved different influences in the experimental and treatment groups: in the experimental group post-intervention, A was more important as a determinant for growth mindset responding. This means that the change in the level of growth mindset likely cannot be interpreted to be due to changes in growth mindset itself. If MI means the same things are measured in different groups, it is not possible for a latent variable whose indicators have different ACE variances to display MI in realistic scenarios. MI will be unattainable unless A, C, and E all have qualitatively indistinguishable effects on responding and the intervention reduced one variance component as much as it increased another one. The first possibility is unlikely. In what world would a genetic effect strongly resemble (when power is moderate) or be identical to (when power is high) the effect of the shared environment? One potential answer is a world in which an A-, C-, and E-influenced common pathway model for a latent variable fully explains the variance in its indicators with and without an intervention (Franić et al., 2013). That does not seem to be our world, as the intervention only affected the mindset measure and not the self-determination composite. Since that is not our world, we know that the intervention must have at least had its effect on the intercept or the residual variance[6] in the mindset indicator, if not both parameters and more. Since the A and C variances definitionally cannot resemble the E variance because it is unsystematic by design, it is irrelevant. Therefore, in scenarios where an AE model (i.e., additive genetics and nonshared environments without shared environmental influence) is strongly confirmed and an intervention affects one or both of those variances, MI must necessarily be violated.

There are certainly more possibilities with biometric variance changes that can be imagined which might be compatible with MI, but they will generally be unfalsifiable, as they'll involve theorizing about quantitatively distinguishable levels of influences that act on phenotypes through different mechanisms yet are still qualitatively indistinguishable in their manifestations.

These three scenarios provided by Protzko (2022), Schneider et al. (2020), and Burgoyne et al. (2018, 2020) are ones in which, first, a different thing is known to be measured because qualitatively different test instruments were used in different groups; second, a highly-specific effect on a cognitive performance measure was elicited based on a novel source of response variance that is independent from general intelligence; and finally, the variance components

involved in just one of a latent variable's *indicators* were confirmed to have been altered by an intervention.

Analysis and Results

Section 1. Protzko (2022)

The first step in fitting this data is to fit the structural equation model, and to fit it thrice: with the whole sample, with one group, and with another group. The model used for these analyses is illustrated in Figure 1. Table 2 contains the loadings for both groups, the one given the meaning of life questionnaire and the one given the meaning of gavagai questionnaire. The differences in their unstandardized loadings[7] were computed and a *p*-value for that difference is provided in the table's last column. For the free will belief factor, the loadings did not significantly differ (*p*'s between .109 and .999); however, for the meaning of life/gavagai factor, four of the loadings significantly differed (*p*'s between 1.7e-7 and .262). This means that metric invariance *must* fail and model fit indices that fail to detect these differences are failing to capture genuine misspecification. When loadings significantly differ, metric invariance fails; when they do not significantly differ, it will tend to be tenable. The fact that metric invariance must fail does not mean that the difference in loadings is necessarily large, and an approximate fit index should still be expected to change with much more considerable changes. But this is beside the point: clearly, the meaning of life and meaning of gavagai questionnaires must be interpreted differently, which is consistent with them measuring different things. As a result, ω cannot be the same unless the loadings were coincidentally higher and lower to equal degrees across measures, so if that is interpreted as a measure of reliability it will likely indicate measurement that is not in common.

Table 3 provides the fit indices for various levels of invariance.[8] The metric, scalar, strict, homogeneity of latent variances, and homogeneity of latent means steps all showcased noninvariance. In other words, they each provided evidence against the naïve comparability of the meaning of gavagai/life groups. To better understand why MI was so clearly violated, partially MI models had to be fitted. These were fitted at each stage and the freed parameters were kept in subsequent stages. The freed parameters were the most noninvariant parameters, selected incrementally, until the partial model did not provide evidence of noninvariance. Table 4 illustrates these results.

The results of partial MI testing were very clear: the free will beliefs questionnaire results were comparable, but the meaning of gavagai/life questionnaires were not. Metric invariance failed first: three of ten loadings had to be freed, and these were all meaning of gavagai/life loadings. Scalar invariance failed next, and two of ten intercepts had to be freed; they were all meaning of gavagai/life intercepts. Strict invariance then failed as well, and this was the most theoretically

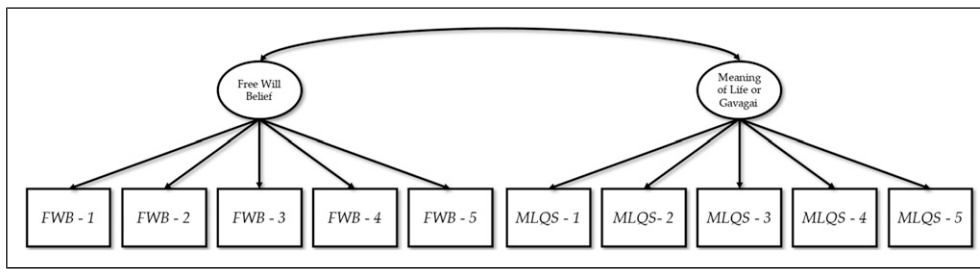

Figure 1. The free will belief—meaning of life/gavagai correlated factor diagram.

Table 2. Baseline Model Factor Loadings and Whether They Significantly Differ.

Factor Variable	Life estimate	Life *SE*	Gavagai Estimate	Gavagai *SE*	*P**
Free will belief					
FWB-1	.981	.045	.878	.046	.109
FWB-2	1.087	.047	1.086	.049	.988
FWB-3	1.211	.049	1.224	.051	.999
FWB-4	1.144	.044	1.172	.048	.999
FWB-5	1.053	.048	1.071	.050	.999
Meaning of life/Gavagai					
MLQS-1	1.344	.049	1.273	.040	.262
MLQS-2	1.536	.048	1.268	.038	1.2e-5
MLQS-3	1.443	.048	1.289	.038	.012
MLQS-4	1.605	.047	1.289	.038	1.7e-7
MLQS-5	1.621	.049	1.339	.040	8.3e-6

*Values of .999 indicate rounding. The configural model was identified by constraining the latent variance to avoid mistakenly constraining noninvariant loadings to equality (Johnson et al., 2009). All results were robust to the use of different identification methods where appropriate. The choice of ML, MLR, or DWLS as the estimator did not qualitatively affect results and loadings were indistinguishable when factors were modeled separately at baseline.

Table 3. Measurement Invariance, Without Partial Models.

Level of invariance	X^2/df	CFI	RMSEA (95% CI)	AIC/BIC
Configural invariance	272.483/68	.983	.063 (.056– .071)	42887/43217
Metric invariance	**302.135/76**	.981	.063 (.056–.071)	**42901**/43187
Scalar invariance	**359.628/84**	.977	.066 (.059–.073)	**42942**/43187
Strict invariance	**1039.201/94**	**.922**	**.116 (.109–.122)**	**43602/43793**
Homogeneity of latent variances	**1063.359/96**	.920	.116 (.110–.122)	**43622/43803**
Equal latent Covariances	1063.360/97	.920	.115 (.109–.122)	43620/43795
Homogeneous latent means	**1175.242/99**	.911	.120 (.114–.127)	**43728/43893**

Note. Bolded values indicate evidence of violations of MI and bolded levels indicate a failure at that level with respect to the previous one. For the column "X^2/df", this means that the difference between a model and the one it is nested under was significant.

informative step, as it is the stage that most-clearly represents a test of the commonality of influences on scores across groups or, in other words, that a questionnaire "measures the same thing" in different groups. The result here was clear, and I have illustrated it in Table 5, which shows that, at each stage of fitting the partially invariant model, the free will belief residual variances made no contribution whatsoever to model fit improvement: the influences on the common questionnaire (free will belief) were shared between groups, but the influences on the uncommon questionnaire (meaning of gavagai/life) were not. The model worked as theoretically expected given that MI is supposed to assess whether a psychological instrument measures the same things in different groups.

Table 4. Measurement Invariance, With Partial Models.

Level of invariance	Freed parameters	X^2/df	CFI	RMSEA (95% CI)	AIC/BIC
Configural invariance	—	272.483/68	.983	.063 (.056–.071)	42887/43217
Metric invariance	**—**	**302.135/76**	.981	.063 (.056–.071)	**42901**/43187
Partial metric invariance	MLQS- 4, 5, and 2 loadings	278.377/73	.983	.061 (.054–.070)	42883/43186
Scalar invariance	As above	**323.664/81**	.980	.063 (.056–.070)	**42912**/43173
Partial scalar invariance	MLQS- 5 and 1 intercepts	285.213/79	.983	.059 (.052–.066)	42878/43149
Strict invariance	As above	**965.745/89**	**.928**	**.115 (.108–.121)**	**43538/43756**
Partial strict invariance	All MLQS residual variances	286.555/84	.983	.057 (.050–.064)	42869/43114
Homogeneity of latent variances	As above	290.720/86	.983	.056 (.049–.064)	42869/43103
Equal latent Covariances	As above	291.036/87	.983	.056 (.049–.063)	42868/43096
Homogeneity of latent means	As above	**413.526/89**	**.973**	.070 (.063–.077)	**42986/43204**
Homogeneous free will mean	Free will mean	291.264/88	.983	.055 (.049–.063)	42866/43089
Homogeneous meaning mean	Meaning of life/ Gavagai mean	**403.427/88**	.974	.069 (.062–.076)	**42978/43201**

Note. Bolded values indicate evidence of violations of MI and bolded levels indicate a failure at that level with respect to the previous one. For the column "X^2/df," this means that the difference between a model and the one it is nested under was significant.

Table 5. Steps of Partial Strict Model Fitting.

Variable	X^2 – Step 1	X^2 – Step 2	X^2 – Step 3	X^2 – Step 4	X^2 – Step 5
Free will belief					
FWB-1	965.743	710.619	552.225	409.111	332.379
FWB-2	965.725	710.603	552.209	409.092	332.362
FWB-3	965.626	710.506	552.116	409.000	332.266
FWB-4	964.533	709.402	551.014	407.895	331.161
FWB-5	965.731	710.608	552.215	409.101	332.369
Meaning of life/Gavagai					
MLQS-1	793.560	**552.227**	—	—	—
MLQS-2	807.884	583.288	**409.113**	—	—
MLQS-3	**710.620**	—	—	—	—
MLQS-4	857.923	634.909	488.060	358.758	**286.555**
MLQS-5	829.249	605.429	455.911	**332.381**	—

Note. Values are X^2 for the model if that parameter was freed. Bolded values indicate a parameter was freed which, accompanied by another stage of parameter inspection, meant that the model continued to fit significantly worse than the partial scalar model. Lower X^2 values indicate better fit, as the model is closer to nonsignificantly differing from the previous one.

Section II. Schneider et al. (2020, I)

Due to the limited number of indicators and the immense support for the concept of psychometric *g* in the intelligence research literature (Savi et al., 2019), a single-factor model was used. Table 6 shows the results of testing for MI in Schneider et al.'s (2020) Experiment I with this model. As that table shows, there was certainly noninvariance in the intercepts for the two figural matrices items, but only one of the items showed a significantly different residual variance, and that effect was only slightly below a *p*-value of .05. In other words, the experimental effect must have been largely homogeneous, or the study had too little power to detect treatment effect heterogeneity.

In terms of magnitude, the experimental group scored 1.30 *g* higher on one figural matrices indicator and 1.15 *g* higher on the other. Using the bias effect size introduced by Lasker and McNaughtan (2022), the amount of this difference between experimental groups that was attributable to bias could be computed. This effect size is a simple extension of the effect size SDI_2 introduced by Gunn et al. (2020) and it can be interpreted in terms of Glass' delta. Briefly, SDI_2 is a signed effect size for measuring the impact of noninvariance in continuous outcomes in the framework of an MGCFA, giving by

$$SDI_2 = \frac{1}{SD_{j2}} \int \left(\widehat{Y}_{j1} - \widehat{Y}_{j2} | \eta \right) \cdot f_2(\eta) d\eta$$

where $\widehat{Y}_{j1} - \widehat{Y}_{j2}$ is the difference in expected scores in different groups (1, 2) for an indicator given some value of the latent variable, η, $f_2(\eta)$ is the second group's distribution of that latent variable, and SD_{j2} is the standard deviation for the second group's scores for the *jth* variable. UDI_2 is the unsigned version of this effect size. This effect size computes the correct magnitude of bias by taking the absolute value of the difference in expected scores in different groups instead of the raw difference between them.

Table 6. Measurement Invariance, With Partial Models.

Level of invariance	Freed parameters	X^2/df	CFI	RMSEA (95% CI)	AIC/BIC
Configural invariance	—	34.865/18	.926	.129 (.062–.193)	4041/4139
Metric invariance	—	40.156/23	.925	.115 (.051–.174)	4037/4121
Scalar invariance	**—**	**103.48/28**	**.668**	**.219 (.175–.265)**	**4090/4161**
Partial scalar invariance	Both figural matrices intercepts	41.69/26	.931	.104 (.036–.160)	4032/4108
Strict invariance	As above	**54.32/32**	**.902**	.112 (.057–.161)	4033/4092
Partial strict invariance	Distraction matrices residual variance	46.94/31	.930	.096 (.028–.149)	4027/4090
Homogeneity of latent variance	As above	47.51/32	.932	.093 (.023–.146)	4026/4086
Homogeneity of latent mean	As above	48.44/33	.932	.091 (.020–.144)	4025/4082

Note. Bolded values indicate evidence of violations of MI and bolded levels indicate a failure at that level with respect to the previous one. For the column "X^2/df," this means that the difference between a model and the one it is nested under was significant.

$$\mathrm{UDI}_2 = \frac{1}{SD_{j2}} \int \left| \widehat{Y}_{j1} - \widehat{Y}_{j2} | \eta | \cdot f_2(\eta) d\eta \right.$$

These effect sizes can be combined to produce an unsigned effect size with the correct magnitude, in the form of SUDI_2, like so:

$$\mathrm{SUDI}_2 = \begin{cases} \mathrm{UDI}_2 \text{if } \mathrm{SDI}_2 \geq 0 \\ \mathrm{UDI}_2 \times -1 \text{if } \mathrm{SDI}_2 < 0 \end{cases}$$

For further information about SDI_2 and UDI_2, see Gunn et al. (2020). To calculate the proportion of an observed gap due to bias with this metric, SUDI_2 can be subtracted from the observed gap if the order of the groups for both effect sizes was consistent.

When SUDI_2 is computed from this model of Schneider et al.'s (2020) data, the greater intercept in the experimental group appears to raise that group's observed scores for either indicator by 1.27 *g* and 1.17 *g*, respectively. Although there is some imprecision in these estimates, it is very likely that the entirety of the difference between the randomized treatment and control groups in figural matrices performance was due to psychometric bias and noise.

Section III. Schneider et al. (2020, II)

Table 7 shows the results of testing for MI in Schneider et al.'s (2020) Experiment II with another single-factor model. The intercept for this experiment's lone figural matrix indicator was significantly noninvariant. The gap between the experimental and treatment groups was 1.21 *g* and the amount of the gap attributable to bias was estimated at 1.08 *g*. It's likely that psychometric bias and noise explain the difference between groups in figural matrix performance in this study, just as in Experiment I.

Section IV. Burgoyne et al. (2020)

The three-indicator self-determination model suggested in Burgoyne et al. (2018) fit their 2020 data perfectly. Because they provided pre- and post-intervention measurements for the treatment and control groups, there were three readily-available means of testing for MI. First,

Table 7. Measurement Invariance, With Partial Model.

Level of invariance	Freed parameters	X^2/df	CFI	RMSEA (95%CI)	AIC/BIC
Configural invariance	—	4.911/4	.993	.045 (0–.154)	5934/6016
Metric invariance	—	9.619/7	.979	.057 (0–.137)	5933/6005
Scalar invariance	—	**61.008/10**	**.584**	**.211 (.162–.263)**	**5978/6040**
Partial scalar invariance	Figural matrix intercept	9.632/9	.995	.025 (0–.110)	5929/5994
Strict invariance	As above	17.320/13	**.965**	**.054 (0–.114)**	5928/5980
Homogeneity of latent variance	As above	17.610/14	.971	.047 (0–.108)	5927/5975
Homogeneity of latent mean	As above	17.771/15	.977	.040 (0–.101)	5925/5970

Note. Bolded values indicate evidence of violations of MI and bolded levels indicate a failure at that level with respect to the previous one. For the column "X^2/df," this means that the difference between a model and the one it is nested under was significant.

Table 8. Measurement Invariance, Prior to Intervention.

Level of invariance	Freed parameters	X^2/df	CFI	RMSEA (95% CI)
Configural invariance	—	0/0	1	0
Metric invariance	—	1.068/2	1	0 (0–.057)
Scalar invariance	—	1.606/4	1	0 (0–.032)
Strict invariance	—	6.770/7	1	0 (0–.042)
Homogeneity of latent variance	—	6.845/8	1	0 (0–.036)
Homogeneity of latent mean	—	7.548/9	1	0 (0–.034)

Note. Bolded values indicate evidence of violations of MI and bolded levels indicate a failure at that level with respect to the previous one. For the column "X^2/df," this means that the difference between a model and the one it is nested under was significant.

Table 9. Measurement Invariance, Prior to Intervention.

Level of invariance	Freed parameters	X^2/df	CFI	RMSEA (95% CI)
Configural invariance	—	0/0	1	0
Metric invariance	—	1.910/2	1	0 (0–.125)
Scalar invariance	—	4.322/4	1	**.018 (0–.100)**
Strict invariance	—	6.186/7	1	0 (0–.073)
Homogeneity of latent variance	—	8.781/8	1	**.020 (0–.080)**
Homogeneity of latent mean	—	8.886/9	1	0 (0–.071)

Note. Bolded values indicate evidence of violations of MI and bolded levels indicate a failure at that level with respect to the previous one. For the column "X^2/df," this means that the difference between a model and the one it is nested under was significant.

testing the MI of self-determination between the control and treatment groups in the pre- and post-intervention timepoints acts as a test of whether the treatment effect results in any amount of bias. Second, testing the MI of self-determination between the control and treatment groups with pre- and post-measurements modeled simultaneously allows for testing whether MI is violated with stable influences controlled. This allows for more powerful estimation of a heterogeneous treatment effect through controlling away stable influences. Finally, testing for longitudinal invariance within both groups lets us know whether the indicators meant the same things within each group, between measurement occasions. If the treatment effect is biasing, the treatment group should see mindset biased between measurement occasions. If Burgoyne et al. (2020) could maintain their sample's attention between measurements, then the control group should not show any such bias, and if there is bias, it should be as likely to show up for any indicator as for the mindset indicator. These models were also used in Section V and the diagonally-weighted least squares estimator was used for both studies.

To save space, most of the tables associated with the tests using Burgoyne et al.'s data were relegated to the supplement. Table 8 shows the result for the pre-treatment MI test, while Table S3 shows the result for the post-treatment MI test. Table S4 shows the results for the combined pre- and post-treatment MI test. Tables S5 and S6 show the results for the control and treatment group's respective longitudinal MI tests. Across each of the tests, MI was attainable exactly as expected: the mindset measure was biased regardless of the method used.

Section V. Burgoyne et al. (2018)

Exactly as before: Tables 9 shows MI testing results for the pre-treatment, and Table S7 show the results of the post-treatment MI test. Table S8 shows the results for the combined pre- and post-treatment MI test. Tables S9 and S10 show the results for the control and treatment group's respective longitudinal MI tests. Across each of the tests, MI was largely attainable as expected, but in the longitudinal test for MI in the treatment group, the effect on the intercepts was only marginally significant ($p = .063$). Given the lower power of this study compared to the earlier one, and the fact that it was an online, MTurk-based study, perhaps a weaker biasing effect shouldn't be too surprising.

Conclusion

MI testing is capable of correctly showing when different things are measured in different groups.

In a reassessment of Protzko (2022), the meaning of life questionnaire was not interpreted the same way as the meaning of gavagai questionnaire. Moreover, the influences on both questionnaires were clearly not wholly shared. The reasons these results are discrepant from those reported with the same data by Protzko (2022) are straightforward. First, Protzko (2022) used insensitive model fit indices that failed to indicate measurement noninvariance where it was tested. Some of this noninvariance had to be apparent on configural model inspection because it was impossible to achieve a different result due to the significance of the loading differences. That it did not appear as indicated with liberal CFI and RMSEA cutoffs indicts those approximate fit indices more than it supports metric or scalar invariance.

Second, Protzko (2022) never fully tested whether the meaning of gavagai/life questionnaires measured the same things in the first place, because strict invariance was not tested. This model must be tested to assess the claim that the same things are being measured because the residual variances represent the influence of *things,* and the equality of construct reliability is also left unknown. The metric and scalar models test the equivalence of the interpretation of certain model parameters and allow researchers to claim that the latent construct of interest is measured at least partially in common while usually allowing the means of latent variables to be legitimately compared, but these steps do not provide information about the reliability of constructs in different groups, nor do they ensure equal reliability of the indicators, or that the influences on scoring are fully shared. Accordingly, in many scenarios, the proposition that the same thing is measured in different groups is likely to be left untested if the strict model isn't fitted.

The interpretation of the results of Protzko (2022) as evidence against one of the central functions of MI testing was always spurious. What the experiment would indicate in the negative (i.e., finding MI holds) is that there is a construct validity problem for the meaning of life and gavagai questionnaires because, whatever they are, they would be measuring the same things! In the affirmative (i.e., finding MI does not hold, whatever the reason), it affirms that there are differences in the involved constructs and that the procedure is able to pick up on them. Theoretically, this test was never able to affirm the claims made in Protzko (2022), only to repudiate them.

To confirm what that study wished to is simple enough. One appropriate design would be to provide the same questionnaire to two groups and bias the results in a clear and visible way, such as by giving out answers to or letting people practice for an objective test instrument, or encouraging one group to answer a certain way on a personality questionnaire while the other is left to answer normally. This has been done in the case of stereotype threat, where a sample was primed to feel threat that reduced their performance on an intelligence test. However, MI testing was also sensitive in that case and it was observed that stereotype threat led to biased measurement

(i.e., noninvariance; Wicherts et al., 2005). In this paper's testing with Schneider et al.'s (2020) figural matrix rule-teaching experiment, it was found that bias was observed for the correct parameters and in practically the exact quantities expected. In a further scenario involving a growth mindset intervention, Burgoyne et al.'s (2018, 2020) datasets showed that growth mindset was biased, as one would expect given their biometric results.

Maassen et al. (2023) reviewed the state of MI testing in the journals *Psychological Science* and *PLOS ONE*. Out of thirteen experimental within-group comparisons, eight were simply noninvariant and five showed supported for scalar variance. There were a further 128 experimental between-group comparisons, of which 67 were simply noninvariant, twelve only reached configural invariance, thirteen reached metric invariance, and 36 reached scalar invariance. In their codebook, forty MI tests were listed as being empirical, with experimental groups, where the groups reached scalar invariance. The characteristics of these studies are worth noting.

Van Dessel and De Houwer (2019) had three comparisons that reached scalar MI: a comparison over time of people's stimulus evaluations, which very significantly changed over time, and two comparisons involving evaluations in hypnosis and relaxation conditions. The lack of any apparent difference in measurement parameters between these two conditions is not very remarkable because there was only one statistically significant difference in evaluations between the conditions at the different timepoints, and it was marginal (two-tailed $p = .018$). Or, in other words, the choice of condition had no apparent effect, so it is unsurprising that measurement wasn't compromised. Luttrell et al. (2019) had a similar result: measurement of persuasiveness were robust to nonsignificant or extremely marginal interactions on the basis of the type of appeal individuals were exposed to. The results of Zlatev (2019) were quite different. In that study, integrity ratings were comparable between conditions where the target was shown as high or low on caring and when participants and targets agreed or disagreed. There were significant differences in integrity ratings by condition, but the meaning of ratings was MI. A handful of other studies reached scalar MI (Ackerman et al., 2018; Berman et al., 2018; Catapano et al., 2019; Kardas & O'Brien, 2018; Moon et al., 2018; O'Brien & Kassirer, 2019; O'Connor & Cheema, 2018; Sawaoka & Monin, 2018; Srna et al., 2018); their common characteristics seemed to be indicating the robustness of perceptions, ratings, etc. to conditions and MI failing to be violated by effects that were any of nonsignificant, marginal, or otherwise very small.

None of those studies showed MI changes in cognitive capabilities, nor a single instance like in Protzko (2022) or Burgoyne et al. (2018, 2020) where MI is implied or practically implied to be violated and it nevertheless appeared to be confirmed. Those results are reassuring about the possibility of doing psychological research involving attitudes and perceptual ratings without bias in measurement, but they otherwise reveal very little about the theoretical import of MI. Nonetheless, the more revealing finding from Maassen et al. (2023) was the number of studies where MI was *not* satisfied. As they showed, even when there's no obvious way that an experiment should cause MI to be violated, it can still be violated, so we cannot take MI for granted and it must always be tested if instruments are going to be interpreted in a way that requires they be unbiased.

At least two other trials have supported the possibility that interventions or formatting differences are detectable through testing MI, one clearly, and one potentially. The clear demonstration came from Arendasy and Sommer (2013). They used a higher order model with crystallized, fluid, and quantitative group factors and showed that variations in the rules involved in figural matrices led to violations of MI in the loadings, intercepts, and residual variances of their figural matrices tasks. Becker et al. (2016) followed similar procedures, but their results were ambiguous because they used conceptually nonsensical models where intelligence, matrices performance, and working memory were separate and correlated group factors, and they ultimately could not all be modeled alongside one another due to convergence problems and they did not adequately test strict invariance in their more limited models. The reanalysis of this data with a

model like Arendasy and Sommer's (2013) would serve as a further test of whether MI means the same things are measured in different groups.

A potential broader implication of these findings is that because interventions aimed at changing traits like growth mindset, spatial ability, or other traits of interest to psychologists usually involve novel sources of potential trait variance, when they cause changes, those changes will tend to be attributable to the resulting noninvariance. As a result, intervention-induced changes in those traits are unlikely to be mediators of the potential beneficial effects of those interventions. In other words, if a causally efficacious trait (e.g., non-cognitive skill, conscientiousness, and grit) predicts success in life and individuals exposed to a trait-boosting intervention become more successful, if the change in the trait was due to psychometric bias, it's unlikely that change is why the intervention caused people to become more successful. Changes in the variable may be indicate the effects of the intervention, but they are unlikely to have the same implications as cross-sectional variance in that trait and we cannot necessarily generalize the effects of the trait observed in the cross-section to those that follow the trial.

The results provided here make for a clear conclusion: a well-powered finding of MI allows psychometricians to claim that the same constructs are measured in different groups.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Jordan Lasker https://orcid.org/0000-0002-5143-2191

Data Availability

All code required to replicate the initial reanalyses is available online at https://rpubs.com/JLLJ/PMI22. The same link provides code for a supplementary analysis of randomly generated Likert data that further illustrates issues with commonly used fit indices and the level of testing in this study's referent. Code for $RMSEA_D$ is at https://rpubs.com/JLLJ/SBFNMC. Code for the reanalysis of Schneider et al.'s (2020) results is located at https://rpubs.com/JLLJ/MatMI. Code for the reanalysis of Burgoyne et al.'s (2018; 2020) results is available at https://rpubs.com/JLLJ/GMMI.

Supplemental Material

Supplemental material for this article is available online.

Notes

1. These quantities are amenable to various interpretations. For the purposes of showing published examples of what researchers understand the different stages of MI testing to mean, I have performed a miniature literature review and placed the results in Table S1 of the supplementary materials (see also Riordan & Vandenberg, 1994).
2. Excluding variables whose influence is totally mediated by latent variables, or whose effects are consistent with as much.
3. This procedure is sometimes reordered.

4. Commentary on this manuscript refers to Version 1 posted on the preprint server PsyArXiv. Version 2 purportedly, but not actually, responds to the model fit criticisms I have made here.
5. One method that may help to escape the conundrum elicited by the arbitrary nature of model fit index cutoffs is to compute dynamic fit cutoffs (i.e., custom simulation-based cutoffs) for the approximate fit indices rather than relying on popular cutoff values that may or may not apply to a given dataset and model (McNeish & Wolf, 2020).
6. The power to detect differences in variances is smaller than the ability to detect differences in means. As such, even though we know the variance in the intervention group was greater due to A expanding while C and E were left unaffected, a significant difference in the residual variances might not be detected for power reasons even though it should theoretically show up.
7. The unstandardized loadings are what affect model fit. Using the standardized loadings, all of them significantly differed for the meaning of life/gavagai factor, but ultimately, only three needed to be freed for partial metric model fitting.
8. Considering the problems with the RMSEA and CFI indices, where one needs reformed (ΔRMSEA) and the other is so problematic as to make its use more onerous than its potential benefits in light of alternatives (ΔCFI), unlike ΔRMSEA, the use of $RMSEA_D$ disconfirms the findings of Protzko (2022), either in terms of typical ΔRMSEA cutoffs (i.e., ΔRMSEA >.015) or confidence intervals for $RMSEA_D$. $RMSEA_D$ is defined in Savalei et al. (2021) as the direct generalization of RMSEA given by the formula $RMSEA_D = \sqrt{\frac{D-df_D}{df_D(N-1)}}$, where RMSEA is $\sqrt{\frac{T-df}{df(N-1)}}$, T is the chi-square test statistic, df is the degrees of freedom, N is the sample size, and D is the model chi-square difference statistic.

References

Ackerman, J. M., Tybur, J. M., & Mortensen, C. R. (2018). Infectious disease and imperfections of self-image. *Psychological Science*, *29*(2), 228–241. https://doi.org/10.1177/0956797617733829

Arendasy, M. E., & Sommer, M. (2013). Reducing response elimination strategies enhances the construct validity of figural matrices. *Intelligence*, *41*(4), 234–243. https://doi.org/10.1016/j.intell.2013.03.006

Beaujean, A. A. (2014). *Latent variable modeling using R* (1st ed.). Routledge.

Becker, N., Schmitz, F., Falk, A. M., Feldbrügge, J., Recktenwald, D. R., Wilhelm, O., Preckel, F., & Spinath, F. M. (2016). Preventing response elimination strategies improves the convergent validity of figural matrices. *Journal of Intelligence*, *4*(1). Article 1. https://doi.org/10.3390/jintelligence4010002

Berman, J. Z., Barasch, A., Levine, E. E., & Small, D. A. (2018). Impediments to effective altruism: The role of subjective preferences in charitable giving. *Psychological Science*, *29*(5), 834–844. https://doi.org/10.1177/0956797617747648

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*(5), 713–734. https://doi.org/10.1177/0022022117749042

Burgoyne, A. P., Carroll, S., Clark, D. A., Hambrick, D. Z., Plaisance, K. S., Klump, K. L., & Burt, S. A. (2020). Can a brief intervention alter genetic and environmental influences on psychological traits? An experimental behavioral genetics approach. *Learning and Motivation*, *72*, 101683. https://doi.org/10.1016/j.lmot.2020.101683

Burgoyne, A. P., Hambrick, D. Z., Moser, J. S., & Burt, S. A. (2018). Analysis of a mindset intervention. *Journal of Research in Personality*, *77*, 21–30. https://doi.org/10.1016/j.jrp.2018.09.004

Catapano, R., Tormala, Z. L., & Rucker, D. D. (2019). Perspective taking and self-persuasion: Why "putting yourself in their shoes" reduces openness to attitude change. *Psychological Science*, *30*(3), 424–435. https://doi.org/10.1177/0956797618822697

Cronbach, L. J. (1947). Test "reliability": Its meaning and determination. *Psychometrika*, *12*(1), 1–16. https://doi.org/10.1007/BF02289289

DeShon, R. P. (2004). Measures are not invariant across groups without error variance homogeneity. *Psychology Science*, *46*(1), 137–149.

Fischer, R., & Karl, J. A. (2019). A primer to (Cross-Cultural) multi-group invariance testing possibilities in R. *Frontiers in Psychology*, *10*, 13. https://doi.org/10.3389/fpsyg.2019.01507

Franić, S., Dolan, C. V., Borsboom, D., Hudziak, J. J., van Beijsterveldt, C. E. M., & Boomsma, D. I. (2013). Can genetics help psychometrics? Improving dimensionality assessment through genetic factor modeling. *Psychological Methods*, *18*(3), 406–433. https://doi.org/10.1037/a0032755

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: A Multidisciplinary Journal*, *27*(4), 503–514. https://doi.org/10.1080/10705511.2019.1689507

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*(4), 466–483. https://doi.org/10.1177/1523422319870726

Johnson, E. C., Meade, A. W., & DuVernet, A. M. (2009). The role of referent indicators in tests of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(4), 642–657. https://doi.org/10.1080/10705510903206014

Kardas, M., & O'Brien, E. (2018). Easier seen than done: Merely watching others perform can foster an illusion of skill acquisition. *Psychological Science*, *29*(4), 521–536. https://doi.org/10.1177/0956797617740646

Lacko, D., Čeněk, J., Točík, J., Avsec, A., Đorđević, V., Genc, A., Haka, F., Šakotić-Kurbalija, J., Mohorić, T., Neziri, I., & Subotić, S. (2022). The necessity of testing measurement invariance in cross-cultural research: Potential bias in cross-cultural comparisons with individualism– collectivism self-report scales. *Cross-Cultural Research*, *56*(2–3), 228–267. https://doi.org/10.1177/10693971211068971

Lasker, J. (2021). Interpreting cross-cultural bias in psychological assessments: An empirical example. PsyArXiv. https://doi.org/10.31234/osf.io/zwb4c

Lasker, J., Haltigan, J. D., & Richardson, G. B. (2022). Measurement issues in tests of the socioecological complexity hypothesis. *Evolutionary Psychological Science*, *8*(2), 228–239. https://doi.org/10.1007/s40806-021-00301-0

Lasker, J., & McNaughtan, J. (2022). Similarities and differences in the structure and interpretation of empowerment and job satisfaction between minority and majority faculty members. *International Journal of Education Policy and Leadership*, *18*(2). Article 2. https://doi.org/10.22230/ijepl.2022v18n2a1249

Little, T. D., Card, N. A., Slegers, D. W., & Ledford, E. C. (2007). Representing contextual effects in multiple-group MACS models. In T. D. Little, J. A. Bovaird, & N. A. Card (Eds.), *Modeling ecological and contextual effects in longitudinal studies* (pp. 121–147). Lawrence Erlbaum Associates Publishers.

Lubke, G. H., & Dolan, C. V. (2003). Can unequal residual variances across groups mask differences in residual means in the common factor model? *Structural Equation Modeling: A Multidisciplinary Journal*, *10*(2), 175–192. https://doi.org/10.1207/S15328007SEM1002_1

Lubke, G. H., Dolan, C. V., Kelderman, H., & Mellenbergh, G. J. (2003). Weak measurement invariance with respect to unmeasured variables: An implication of strict factorial invariance. *British Journal of Mathematical and Statistical Psychology*, *56*(2), 231–248. https://doi.org/10.1348/000711003770480020

Luttrell, A., Philipp-Muller, A., & Petty, R. E. (2019). Challenging moral attitudes with moral messages. *Psychological Science*, *30*(8), 1136–1150. https://doi.org/10.1177/0956797619854706

Maassen, E., D'Urso, E. D., Van Assen, M. A. L. M., Nuijten, M. B., De Roover, K., & Wicherts, J. M. (2023). The dire disregard of measurement invariance testing in psychological science. *Psychological Methods*. Online ahead of print. https://doi.org/10.1037/met0000624

McNeish, D., & Wolf, M. G. (2020). Dynamic fit index cutoffs for confirmatory factor analysis models. *PsyArXiv*. https://doi.org/10.31234/osf.io/v8yru

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

Meuleman, B., Żółtak, T., Pokropek, A., Davidov, E., Muthén, B., Oberski, D. L., Billiet, J., & Schmidt, P. (2022). Why Measurement Invariance is Important in Comparative Research. A Response to Welzel et al. (2021). *Sociological Methods & Research*, Online ahead of print. https://doi.org/10.1177/00491241221091755

Millsap, R. E. (2011). *Statistical approaches to measurement invariance* (1st ed.). Routledge.

Moon, J. W., Krems, J. A., & Cohen, A. B. (2018). Religious people are trusted because they are viewed as slow life-history strategists. *Psychological Science*, *29*(6), 947–960. https://doi.org/10.1177/0956797617753606

O'Brien, E., & Kassirer, S. (2019). People are slow to adapt to the warm glow of giving. *Psychological Science*, *30*(2), 193–204. https://doi.org/10.1177/0956797618814145

O'Connor, K., & Cheema, A. (2018). Do evaluations rise with experience? *Psychological Science*, *29*(5), 779–790. https://doi.org/10.1177/0956797617744517

Penfield, R. D., & Camilli, G. (2006). 5 differential item functioning and item bias. In C. R. Rao, & S. Sinharay (Eds.), *Handbook of Statistics*. (*26*, pp. 125–167). Elsevier. https://doi.org/10.1016/S0169-7161(06)26005-X

Plomin, R., DeFries, J. C., McClearn, G. E., & McGuffin, P. (2008). *Behavioral genetics* (5th ed.). Worth Publishers.

Protzko, J. (2018). Null-hacking, a lurking problem. PsyArXiv. https://doi.org/10.31234/osf.io/9y3mp

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review: Developmental Review*, *41*, 71. https://doi.org/10.1016/j.dr.2016.06.004

Riordan, C. M., & Vandenberg, R. J. (1994). A central question in cross-cultural research: Do employees of different cultures interpret work-related measures in an equivalent manner? *Journal of Management*, *20*(3), 643–671. https://doi.org/10.1177/014920639402000307

Ropovik, I. (2015). A cautionary note on testing latent variable models. *Frontiers in Psychology*, *6*, 1715. https://doi.org/10.3389/fpsyg.2015.01715

Savalei, V., Brace, J., & Fouladi, R. T. (2021). We need to change how we compute RMSEA for nested model comparisons in structural equation modeling. *OSF*. https://doi.org/10.31234/osf.io/wprg8

Savi, A. O., Marsman, M., van der Maas, H. L. J., & Maris, G. K. J. (2019). The wiring of intelligence. *Perspectives on Psychological Science*, *14*(6), 1034–1061. https://doi.org/10.1177/1745691619866447

Sawaoka, T., & Monin, B. (2018). The paradox of viral outrage. *Psychological Science*, *29*(10), 1665–1678. https://doi.org/10.1177/0956797618780658

Schneider, B., Becker, N., Krieger, F., Spinath, F. M., & Sparfeldt, J. R. (2020). Teaching the underlying rules of figural matrices in a short video increases test scores. *Intelligence*, *82*. https://doi.org/10.1016/j.intell.2020.101473

Srna, S., Schrift, R. Y., & Zauberman, G. (2018). The illusion of multitasking and its positive effect on performance. *Psychological Science*, *29*(12), 1942–1955. https://doi.org/10.1177/0956797618801013

Svetina, D., Rutkowski, L., & Rutkowski, D. (2020). Multiple-group invariance with categorical outcomes using updated guidelines: An illustration using mplus and the lavaan/semTools packages. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(1), 111–130. https://doi.org/10.1080/10705511.2019.1602776

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*(1), 4–70. https://doi.org/10.1177/109442810031002

van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, *9*(4), 486–492. https://doi.org/10.1080/17405629.2012.686740

Van Dessel, P., & De Houwer, J. (2019). Hypnotic suggestions can induce rapid change in implicit attitudes. *Psychological Science*, *30*(9), 1362–1370. https://doi.org/10.1177/0956797619865183

Welzel, C., Brunkert, L., Kruse, S., & Inglehart, R. F. (2021). Non-invariance? An overstated problem with misconceived causes. *Sociological Methods & Research*, *52*(3), 0049124121995521. https://doi.org/10.1177/0049124121995521

Wicherts, J. M. (2016). The importance of measurement invariance in neurocognitive ability testing. *The Clinical Neuropsychologist*, *30*(7), 1006–1016. https://doi.org/10.1080/13854046.2016.1205136

Wicherts, J. M., Dolan, C. V., & Hessen, D. J. (2005). Stereotype threat and group differences in test performance: A question of measurement invariance. *Journal of Personality and Social Psychology*, *89*(5), 696–716. https://doi.org/10.1037/0022-3514.89.5.696

Zlatev, J. J. (2019). I may not agree with you, but I trust you: Caring about social issues signals integrity. *Psychological Science*, *30*(6), 880–892. https://doi.org/10.1177/0956797619837948