Original Research Article

Educational and Psychological
Measurement
1–25
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/00131644241282982
journals.sagepub.com/home/epm
Sage

Invariance: What Does Measurement Invariance Allow Us to Claim?

John Protzko[1]

Abstract

Measurement involves numerous theoretical and empirical steps—ensuring our measures are operating the same in different groups is one step. Measurement invariance occurs when the factor loadings and item intercepts or thresholds of a scale operate similarly for people at the same level of the latent variable in different groups. This is commonly assumed to mean the scale is measuring the same thing in those groups. Here we test the assumption of extending measurement invariance to mean common measurement by randomly assigning American adults (N = 1500) to fill out scales assessing a coherent factor (search for meaning in life) or a nonsense factor measuring nothing. We find a nonsense scale with items measuring nothing shows strong measurement invariance with the original scale, is reliable, and covaries with other constructs. We show measurement invariance can occur without measurement. Thus, we cannot infer that measurement invariance means one is measuring the same thing, it may be a necessary but not a sufficient condition.

Keywords

psychometrics, latent variable, psychological measurement, measurement invariance

How do we measure what is going on in someone's mind? How do we measure their beliefs, their aspirations, their abilities? Measurement is a complicated process involving the discovery of the mappings between a numerical system (e.g., algebraic relationships) and empirical relationships (e.g., qualitative relationships between phenomena; Luce et al., 2007). Psychological measurement is even more difficult, as the qualitative phenomena and relationships (e.g., mental life) cannot be directly observed. Psychometrics built on latent variable modeling is the current answer to

[1] Central Connecticut State University, New Britain, USA

Corresponding Author:
John Protzko, Central Connecticut State University, New Britain, CT 06050, USA.
Email: protzko@gmail.com

how we measure psychological phenomena (e.g., Borsboom, 2005; Borsboom et al., 2003; Markus & Borsboom, 2013). While there have been challenges to psychometrics as a foundation of psychological measurement (e.g., Michell, 2021; Uher, 2021), our conceptions of validity, reliability, scale construction, are all grounded in psychometric modeling.

When administering the same scale to different groups of people or the same people across time, one needs to be sure the scale is measuring the same construct in all groups or across time. A simple addition scale, with questions like ''what is 4 + 3?'' given to young children just learning to add, tests their ability to apply the rules of arithmetic. That same scale given to adults is more a test of long-term memory for math facts. Thus, the same scale is not measuring the same underlying process in both young children and adults.

To assess whether a scale is assessing the same underlying processes or construct, tests of measurement invariance are applied. Measurement invariance testing is the process of assessing the psychometric properties of a scale in different groups or across time. Testing assesses whether people with the same level of an underlying ability score similarly on all items and whether the items equally reflect the underlying construct. When measurement invariance is found, the conclusion is often that the scale is measuring the same ''thing'' in all groups or across time. Here, we challenge this assumption of measurement invariance meaning the same underlying ability or construct is being measured; by showing measurement, invariance can occur when not consistently measuring anything.

First, we review evidence that finding measurement invariance is commonly interpreted as measuring the same thing across groups or time. Then, we highlight an important but undervalued study in the history of measurement invariance suggesting measurement invariance can be found when the same processes are quite different. Next, we directly test the assumption of common measurement from measurement invariance in a large preregistered study of American adults. Finally, we briefly discuss the conclusions of our argument for both the valuable practice of measurement invariance and the implications for psychometric measurement.

What is Measurement Invariance?

Measurement invariance is the investigation of the psychometric properties of a multi-item scale either in different groups of participants (Jöreskog, 1971) or in the same group of participants over time. The process of testing measurement invariance in the multigroup case (the more relevant to our purposes here) is as follows: (1) administer the same scale to different groups of individuals under the same conditions. This can be naturally occurring groups (e.g., ethnic groups, cultures, ages; see Dolan et al., 2006, for an example) or groups randomized to conditions (see Pages et al., 2022; Protzko et al., 2019; Whitehurst et al., 1994, for examples). (2) Make sure the same factor structure exists in both groups (e.g., same number of factors with items loading on the same factor[s]); record a measure of model fit (often CFI

& RMSEA). (3) Constrain the factor loadings of all items to the factors to be the same in all groups. There are numerous ways to test whether the invariance restrictions degrade model fit. One method includes difference-based measures such as the change (Δ) in CFI or RMSEA. If ΔCFI or ΔRMSEA is within some threshold (often <.01 when following Cheung & Rensvold, 2002) move to the next step in testing; if the degrade in model fit is higher than the threshold, stop and conclude the items are not reflecting the underlying construct in all groups the same. (4) Constrain the means of the items (or thresholds for categorical items) to be the same in all groups. If the change (Δ) in fit statistics remains below some critical threshold, continue imposing restrictions—otherwise stop and conclude the items are not being answered the same way in all groups for people with the same level of the underlying latent trait.[1]

There are sometimes additional steps of imposing equal errors across all groups, which is desirable but not considered necessary for measurement invariance. One may also switch the order, starting with imposed means and factor loadings and progressively release constraints, testing if model fit (e.g., CFI/RMSEA) *improves* >.01 (see Horn & McArdle, 1992).

Measurement Noninvariance

When any of the above steps fail and model fit decreases too much as a function of constraining the factor loadings or means/thresholds to be equal, measurement invariance does not hold and the scale is said to have measurement noninvariance. Measurement noninvariance can occur when different pressures or histories cause the scale to operate differently across groups or time and is an important indicator of differential item functioning. Teaching to the test, for example, can create measurement noninvariance at the item means or threshold level because individuals with an average level of ability in one group may correctly answer difficult items from memory (opposed to reasoning it out; see Protzko, 2016 for a brief discussion). If an intervention teaches children the answers to matrix reasoning problems, for example, when tested, the matrix problems are no longer simply measuring reasoning ability but also memory. Furthermore, individuals so trained may score higher than others of a comparable reasoning ability on the trained test but not on untrained items from a different test.

Certain manipulations can also induce participants to respond to a scale randomly, or in a specific calculated manner, which can alter the correlations among the items and subsequently, how they load on to the underlying factor (see Protzko et al., 2019 for a brief discussion). Historical changes can also lead to longitudinal measurement noninvariance. In cohorts from the 1990s through the 2010s administered the Narcissistic Personality Inventory, some items showed measurement noninvariance. Asking someone to choose which is more true of them: ‘‘When people compliment me I sometimes get embarrassed’’ or ‘‘I know that I am good because everybody keeps telling me so’’ show noninvariant means and loadings across generations

(Wetzel et al., 2017). This means the ‘‘narcissistic’’ response (‘‘I know that I am good because everybody keeps telling me so’’) was correlating with other narcissistic responses for adults in the 1990s but not correlating the same for adults in the 2010s. This could be because children who grew up in later cohorts were enmeshed in the self-esteem movement of the 80s and heard compliments about themselves to greater extents, hypothetically. Therefore, the item is noninvariant because it is no longer measuring narcissism the same way it used to.

What is Measurement Invariance Taken to Mean?

At its simplest and most descriptive—measurement invariance means that for a multi-item scale with an underlying factor structure, in multiple groups, people of the same ability level across groups are answering the items the same way and the items load onto the same factor the same way. A prevalent interpretation of this statistical state of affairs, however, is that one has evidence that one is measuring the same *thing* in those groups.

This assumption of common measurement from measurement invariance was not always held. The earliest work on measurement invariance (specifically factorial invariance) was to see if the same factor pattern occurs in all groups to conclude whether the groups came from the same population (e.g., Meredith, 1964a, 1964b). In the late 1960s and early 1970s, it began to be questioned whether showing factorial invariance meant one could assign the same substantive interpretation to the factors in all populations. This work drew largely on the administration of intelligence batteries to people from different cultures. If factorial invariance can be shown with the same battery in different cultures, what does that mean?

Where cross-cultural work showed preliminary factorial invariance in intelligence batteries in English, Scottish, Eskimo, Canadian Native, Ugandan, and Jamaican children (Vernon, 1967; later Vernon, 1969); it was questioned whether any interpretation of common measurement could be supported without, at minimum, factorial invariance (Irvine, 1969). This question of whether factorial invariance meant one could assign the same substantive interpretation was also investigated in batteries given to people of different socioeconomic statuses (McGraw & Jöreskog, 1971). This interpretation was kept up in cross-cultural research of factorial invariance as common measurement (see also Buss & Royce, 1975). In the late 1970s, measurement invariance was being touted as being evidence for a trait measuring the same thing (Rock et al., 1978). The shift in interpretation from common population to common measurement seems to have taken hold by the 1980s with work testing whether factorial invariance held between adults and the elderly meant testing whether psychometric measures measure the same construct (e.g., Cunningham, 1980).

The conclusion that people have been interpreting measurement invariance is as indicative of a test measuring the same thing in groups or across time is not an extrapolation, but comes directly from numerous studies on measurement invariance. To demonstrate examples of this interpretation, we include direct quotes from a sampling of articles on measurement invariance:

> Measurement invariance is the equivalence of a measured construct in two or more groups, such as people from different cultures. It assures that the same constructs are being assessed in each group. (Chen, 2008, p. 1005)
>
> Any comparison of the same constructs across time or across groups assumes that the measurements are factorially invariant. (Little, 2013, p. 137)
>
> Tests of measurement invariance address . . .Is the same construct measured in all groups?'' (Steinmetz et al., 2009, p. 600)
>
> Factorial invariance within longitudinal structural equation models: Measuring the same construct across time'' (Widaman et al., 2010, p. 10 (Title))
>
> . . .we are measuring the same thing in the same metric at each occasion. Rather than assuming this to be the case, we can formally test this hypothesis of factorial invariance, but only if multiple indicators of a construct are available at each occasion. (Widaman et al., 2010, p. 11)
>
> we must determine that the same constructs are being assessed in each group. (Bowers et al., 2010, p. 723)
>
> measurement invariance is conducted to determine whether the items and the underlying dimensions mean the same thing across the 12 countries. (Kim et al., 2013, p. 83)
>
> Noninvariance can be informative and may lead researchers to important conclusions about how different groups interpret the same construct. (Putnick & Bornstein, 2016, p. 87)
>
> Measurement invariance refers to whether the scale functions equivalently for participants from different cohorts. Put simply, this means that the same latent trait is being measured in the same way across cohorts. (Wetzel et al., 2018, p. 135).
>
> It is seen, thus, to make the case that one is measuring the same thing in different samples of subjects or at different times for the same subjects, one could test an hypothesis stipulating that the columns of the factor pattern are invariant across the different conditions represented by different samplings of subjects. (Horn & McArdle, 1992, p. 117)

It is thus no small claim that a prevalent interpretation of measurement invariance is that once achieved, one is measuring the same thing in different groups or measuring the same thing across time.

One reviewer also pressed us to provide a comprehensive meta-analysis as to the rate of this interpretation. Instead of a comprehensive meta-analysis, we ran a simple test of drawing a random sample of articles from PsychInfo[2] using the search terms ''measurement invariance'' and loading. This returned 537 articles ever indexed by PsychInfo. We took a random 10% of these articles using a random number generator from random.org, eliminated non-peer-reviewed articles and articles not in a language readable by us. We then read each article and coded if the article repeated the ''invariance means measuring the same thing'' claim. Data, timestamp of random number draw, coding, and quotes from the text where the claim is made are available at https://osf.io/qvfbn/. In short, we find that approximately 45% of articles on measurement invariance repeat this claim that measurement invariance means one is measuring the same thing. So the interpretation is certainly prevalent.

Central Argument

The central argument of our manuscript is as follows: a prevalent interpretation of measurement invariance is if you have measurement invariance, you are measuring the same thing. Through contraposition that statement is logically equivalent to: if you are not measuring the same thing, you will not have measurement invariance (i.e., you will have measurement noninvariance). The interpretation of measurement invariance, that it means you are measuring the same thing in different groups/across time, is what we investigate.

Previous Evidence

Previous evidence against the common measurement interpretation of measurement invariance comes from a paper on mathematical reasoning (Widaman et al., 1992). In this study, 24 second graders and 100 college freshmen were given the same mathematical reasoning scale. This scale consisted of a number of True–False statements about whether an addition problem and answer was right (for example: True or False: 24 + 24 = 48). Even though to solve these problems second graders would have to calculate the values first (using arithmetical reasoning) and college freshmen should only have to check their long-term memory for math facts, measurement invariance was found between the two groups (Widaman et al., 1992).

If one is to believe that measurement invariance means one is capturing the same mental process, this should be surprising; the mathematical reasoning scale is operating, presumably, very differently in second graders and college freshmen. The mental processes giving rise to the responses are potentially entirely separate (note college freshman finished the test considerably faster than second graders; Widaman et al., 1992). It is thus odd that a test *could* be measuring different mental processes in different groups but still show evidence of measurement invariance. Although a small study, it is one piece of earlier evidence that measurement invariance is insufficient to reason that one is measuring the same ''thing.''

Another study tested whether a coherent factor could be extracted from complete nonsense items. The investigators started by adapting the Theories of Intelligence scale (Dweck et al., 1995) which includes items such as ''Your intelligence is something about you that you can't change very much'' and changed the word ''intelligence'' in all items to a nonsense word (gavagai, e.g., ''Your gavagai is something about you that you can't change very much''; Maul, 2017). Participants filling out the ''Theories of Gavagai'' scale showed evidence of a coherent and reliable factor with high loadings and good model fit. This ‘‘measurement’’ (e.g., good model fit on a single factor) was also seen when all words were changed to nonsense words (e.g., ‘‘Lorem ipsum vault dore valdis’’) and when items were completely blank and the participants had to answer blank items on a Likert-type scale (Maul, 2017).

One reason this can happen is that the use of nonsense items may be perceived as a breach of the social contract between the researcher and the respondent (certainly a break in norms of communication, see Grice, 1975). This could result in response styles that may not be reflective of the participant’s true beliefs or attitudes.

Although tests of measurement invariance could not conducted in that study, it is curious that traditional methods of ‘‘measurement’’ (e.g., single well-fitting factor with reliability) could be seen in the absence of anything to be measured.

Here we build on these two previous works by investigating whether measurement invariance can be found between groups of individuals when we know the tests they are given are not measuring the same thing. If measurement invariance can be shown with no common measurement, the statement ‘‘If not measuring the same thing then no measurement invariance’’ is incorrect and its contraposition ‘‘if measurement invariance then you are measuring the same thing’’ is likewise incorrect.

Method

Participants were randomly assigned to take two versions of a scale, one measuring the search for meaning in life subscale (Steger et al., 2006) or an altered nonsense version which changed each word of ‘‘meaning’’ or ‘‘purpose’’ with the term ‘‘gavagai.’’ All items were presented in random order.

> Please take a moment to think about what makes your life and existence feel important and significant to you. Please respond to the following statements as truthfully and accurately as you can, and also please remember that these are very subjective questions and that there are no right or wrong answers. Please answer according to the scale below: Absolutely Untrue, Mostly Untrue, Somewhat Untrue, Can’t say, Somewhat True, Mostly true, Absolutely True [7 point unnumbered scale with response options Absolutely Untrue, Mostly Untrue, Somewhat Untrue, Can’t say, Somewhat True, Mostly true, Absolutely True]
>
> I am looking for something that makes my life feel meaningful.
>
> I am always looking to find my life’s purpose.
>
> I am always searching for something that makes my life feel significant.
>
> I am seeking a purpose or mission for my life.
>
> I am searching for meaning in my life.

The search for Gavagai scale had the following instructions and all items in random order as well:

> Please take a moment to think about gavagai. Please respond to the following statements as truthfully and accurately as you can, and also please remember that these are very subjective questions and that there are no right or wrong answers. Please answer according to the scale below: Absolutely Untrue, Mostly Untrue, Somewhat Untrue, Can't say, Somewhat True, Mostly true, Absolutely True [1-7 unnumbered scale with response options Absolutely Untrue, Mostly Untrue, Somewhat Untrue, Can't say, Somewhat True, Mostly true, Absolutely True]
>
> I am looking for something that makes my life feel gavagai.
>
> I am always looking to find my gavagai.
>
> I am always searching for something that makes my life feel gavagai.
>
> I am seeking a gavagai for my life.
>
> I am searching for gavagai in my life.

Participants also filled out a scale about their belief in Free Will. This was done to test whether the nonsense version of the scale would correlate with the belief in Free Will (unpublished data available here https://osf.io/68ax4/ and here https://osf.io/f46yx/). Participants were randomly assigned to take the Free-Will Beliefs scale or the meaning scale in random order. For the Free-Will Beliefs scale (Nadelhoffer et al., 2014) participants read with items in random order:

Please read the following sentences carefully and then indicate your level of agreement [all on a 7-point unnumbered scale from Strongly disagree, Disagree, Somewhat disagree, Neither agree nor disagree, Somewhat agree, Agree, to Strongly agree].

> People always have the ability to do otherwise.
>
> People always have free will.
>
> How people's lives unfold is completely up to them.
>
> People ultimately have complete control over their decisions and their actions.
>
> People have free will even when their choices are completely limited by external circumstances.

On the final page, participants in the nonsense factor condition were asked what they believed the word ''gavagai'' meant:

> You just answered some questions about Gavagai, what does that mean to you? What do you think gavagai is? [response options in random order except for 'I do not know' which is always at top' [I do not know; meaning/purpose; happiness; money; sexual pleasure].

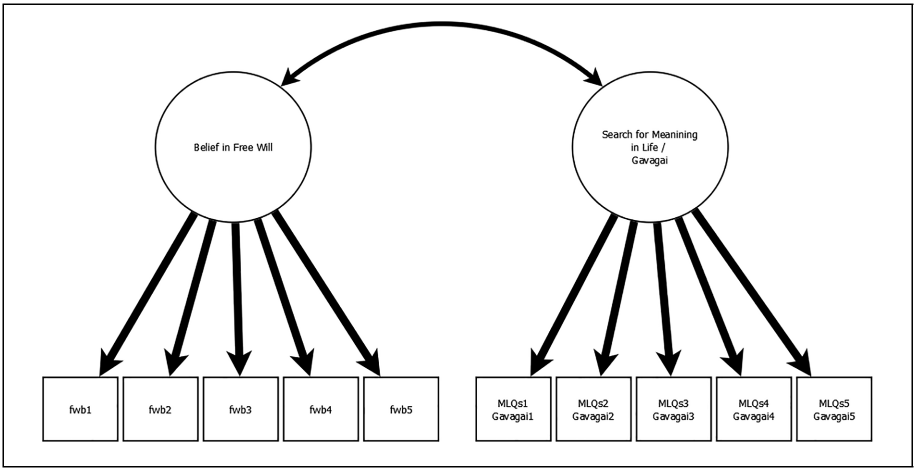

Figure 1. Approach for Testing Measurement Invariance by Randomly Assigning People to Get a Search for Meaning in Life Scale or a Modified Version That Measures Nonsense (e.g., Search for Gavagai).

This was done to test whether participants could intuit the meaning from the nonsense scale, undercutting our measurement investigation test.

Data Analysis

We first constructed a single-factor model for the search for meaning in life subscale (consistent with Steger et al., 2006) and a single-factor model for the Free-Will Beliefs scale (see Figure 1). Our original coding included latent search for meaning predicting latent free-will, which cannot be used in measurement invariance testing (Millsap, 1997). We therefore had to change this to a covariance instead of a causal path. We then tested measurement invariance by first imposing factorial invariance, then adding mean invariance on the search for meaning in life/gavagai factor. We used the cutoff of ΔCFI or ΔRMSEA worsening by <.01 (smaller CFI, larger RMSEA) as our criteria for accepting or rejecting evidence for different levels of measurement invariance (following Cheung & Rensvold, 2002).

Participants

This study was part of a large-scale investigation into the discovery or new findings in the social-behavioral sciences. As such, we used a fixed 1,500 participants drawn from an online panel company in a stratified sample with unequal probabilities of selection so that the people who completed the questionnaire would resemble the nation's adult population (according to the most recently available Current

Table 1. Model Fit Statistics.

Constraint	CFI	ΔCFI	RMSEA	ΔRMSEA
Same factor/configural invariance	.984	.003	.041	−.009
Same loadings/factorial invariance	.982	−.002	.042	+.001
Same loadings + intercepts/scalar invariance	.977	−.005	.047	+.007

Population Survey, conducted by the U.S. Census Bureau) in terms of gender, age, education, ethnicity (Hispanic vs. not), race (allowing each participant to select more than one race), region, and income. This study was preregistered prior to data collection at https://osf.io/5zw83/.

Results

We first imposed the same dual single-factor models for both the belief in Free Will and the Search for Meaning in Life/Gavagai. We set up a multigroup confirmatory factor analysis with one group having the search for meaning factor and the other group having the search for Gavagai factor. The dual single-factor model fit well in both groups (CFI = .981, RMSEA = .05).

Constraining the search for gavagai variables to have the same factor loadings as their respective search for meaning variables did not results in a meaningful loss of model fit (see Table 1). Constraining the search for gavagai variables to have the same means as their respective search for meaning variables, given the same level of latent variable, did not results in a meaningful loss of model fit (Table 1). Finally, constraining the variables to have the same means and factor loadings likewise did not significantly decrease model fit.

For both the search for meaning and the search for Gavagai, the factor loading for the variables was very high (.83–.92). This suggests that a ‘‘factor’’ can be made that does not really measure anything (due to the nonsense words) and it forms a coherent strongly loaded factor (replicating Maul, 2017).

Furthermore, this nonsense factor also correlated significantly with the belief in Free Will. In the search for meaning in life group, latent search for meaning in life correlated significantly and strongly with the belief in free will (r = .42, $p < .001$, 95% confidence interval [CI] = .28–.56). In the search for gavagai group, the correlation was smaller but the search for gavagai still significantly correlated with the belief in free will (r = .3, $p < .001$, 95% CI = .18–.41) with overlapping 95% confidence intervals. This suggests that the correlation between the belief in Free Will and Search for Meaning in Life is largely a function of sharing a common response format, as there is a significant correlation with a nonsense factor as well.

After the study was ran, participants in the gavagai group were asked what they believed the term gavagai meant while they were filling out the survey. Sixteen percent of participants indicated they thought gavagai mean meaning in life. While this

Table 2. Model Fit Statistics for More Restrictive Models Only in Participants Who Did Not Believe Gavagai Meant Meaning in Life.

Constraint	CFI	ΔCFI	RMSEA	ΔRMSEA
Same factor/configural invariance	.984	.003	.041	−.009
Same loadings/factorial invariance	.981	−.003	.042	+.001
Same loadings + intercepts/scalar invariance	.975	−.006	.046	+.004

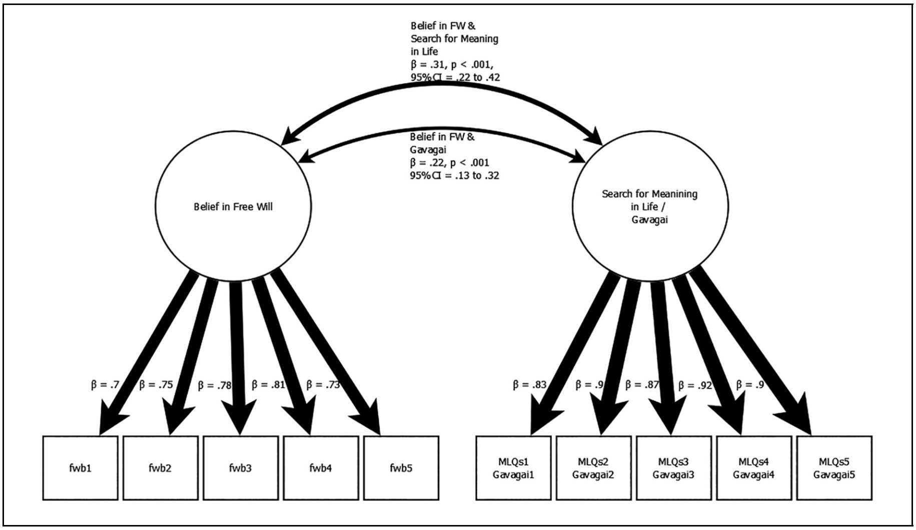

Figure 2. Factor Loadings for the Belief in Free Will (FWB) and Search for Meaning/Search for Gavagai factors (Search).

Factor Loadings are the same for the Search for Meaning in Life factor and the Nonsense factor (via strong measurement invariance). Paths are weighted in the display by their standardized value.

may be an overestimate, as participants may have seen the option and it made the most sense, we eliminated everyone who selected that option and ran the model again.

Testing measurement invariance without anyone who could have intuited the Gavagai scale as meaning in life, we again found evidence for strong measurement invariance. The constraints to model fit did not significantly decrease when constraining the factors to have the same factor loadings, means, or means and factor loadings (see Table 2).

Furthermore, all of the items loaded just as highly on the search for Gavagai factor when participants who guessed correctly were included or excluded (.83 to .92; see Figure 2).

Finally, removing those who guessed the nature of gavagai correctly did decrease the covariance between latent belief in Free Will and the Search for Meaning/Gavagai. In the search for meaning group, these beliefs were strongly correlated (r = .43, p < .001, 95% CI = .57–.29). In the search for Gavagai group, these beliefs were less strongly, though still significantly, correlated (r = .25, p < .001, .37–.14).

Thus, in our preregistered analysis, we were able to create a nonsense factor that had five nonsense items with factor loadings (all > .83) and high reliability (ω = .91). This nonsense factor significantly correlated with the belief in free will, despite not measuring anything. Overall, we achieved measurement invariance when no meaningful measurement occurred.

Exploratory Analyses

A number of commentators have suggested alternate approaches to measurement invariance, especially in the context of this paper. Here we present a number of analyses not part of the original preregistered decisions, briefly presenting possible justifications, and placing into the broader goal of this paper.

Alternate Tests of Measurement Invariance

The original analysis preregistered testing measurement invariance by testing whether model fit decreased as a function of each new constraint, with significant model fit decrement using the cutoff of ΔCFI or ΔRMSEA worsening by more than .01. This was chosen a priori as it is among the most common and supported criteria for testing measurement invariance (Cheung & Rensvold, 2002). It is not the only one.

Additional Metrics of Model Fit

Dynamic Fit Indices. It could be possible that configural invariance is indeed not met. Traditional metrics of model fit (e.g., CFI, RMSEA, SRMR) are not built for single-factor models as they were meant to test for misspecifications like cross-loadings (Hu & Bentler, 1999; as described in McNeish & Wolf, 2022). Dynamic Fit Indices are a method of assessing model fit in the single-factor model by using Monte–Carlo methods to simulate a data set both with and without unmodeled residual covariances (see McNeish & Wolf, 2021).

We tested the data set that removed the people who inferred what ‘‘gavagai’’ meant. Testing the two single-factor models using Dynamic Fit Indices suggests that for the search for meaning in life, model fit is good assuming 1/3 of the items have an unmodeled residual correlation. As can be seen in Table 3, model fit was excellent for both CFI and SRMR. For RMSEA, model fit was worse than the cutoffs for both versions of the scale (see McNeish & Wolf, 2022).

Given the large sample size, large loadings (all > .83), and low number of items, based on the simulations on Dynamic Fit Indices (McNeish & Wolf, 2022), SRMR has the lowest rate of false rejections of a single-factor model and CFI has the highest

Table 3. *Dynamic Fit Indices for the Single-Factor Models in Both the Search for Meaning in Life and the Gavagai Versions of the Scale.*

	Search for meaning in life			Gavagai		
level	SRMR	RMSEA	CFI	SRMR	RMSEA	CFI
Level 1: 95/5	.012	.087[a]	.992	.006	.077[a]	.996
Level 2: 95/5	.018	.13[a]	.983	.009	.131[a]	.989

[a]Indicates instances where model fit is worse than a prespecified cutoff.

Table 4. Measurement Invariance Based on Tucker–Lewis Index (TLI).

Constraint	TLI	ΔTLI
Same factor (configural)	.993	
Same loadings (metric)	.988	–.005
Same loadings + intercepts (scalar)	.979	–.009

Using ΔTLI of $>$ .01 as evidence of measurement invariance, this method of comparing TLI, despite having an inflated rate of type I errors (Sass et al., 2014) shows strong evidence of measurement invariance.

rate of correct rejections of misspecified single-factor models. It is important to note, however, the differences between the indices in the simulations of McNeish & Wolf (N = 400, eight-items, Loadings around .9) are effectively identical (within 1 percentage point). Taken together, we argue that configural invariance using dynamic fit indices is likely met. Thus, this is not a threat to our investigation here.

TLI. The Tucker–Lewis Index (TLI; Tucker & Lewis, 1973) has also seen some use in testing for measurement invariance (e.g., Marsh et al., 2009, 2010). It was not used a priori as previous investigations have shown ΔTLI is correlated with the overall model fit (Cheung & Rensvold, 2002) and proposed cutoffs for TLI when testing measurement invariance perform poorly in simulations by inflating type I error rates (e.g., Sass et al., 2014).

Yet, for completeness, we present the results of our measurement invariance testing using ΔTLI with the cutoff of ΔTLI $<$ .01 for evidence of measurement noninvariance. Results support measurement invariance between the two scales, as adding same factor loading constraints or item intercept constraints or both do not show a reduction in model fit greater than ΔTLI $>$ .01 (see Table 4).

Thus, using the ΔTLI method, we would not see full measurement invariance. Although the measures of model fit are all good, in such a situation one can look for partial invariance or using other methods like alignment models (which we pursue below).

Table 5. Likelihood Ratio Test of χ^2 Values Testing Measurement Invariance.

Constraint	LR data	LR test
Same factor (configural)	Log-likelihood of model: −8,798.211 scaling correction factor: 2.3663 # free parameters: 30	
Same loadings (metric)	Log-likelihood of model: −8,812.349 scaling correction factor: 2.5489 # free parameters: 26	χ^2_{LR} (4) = 28.28/1.18 = 23.97, $p < .001$
Same loadings + intercepts (scalar)	Log-likelihood of model: −8,839.167 scaling correction factor: 2.8116 # free parameters: 22	χ^2_{LR} (4) = 53.64/3.27 = 16.4, $p = .037$

Results indicate the scales are not completely measurement invariant.

χ^2 Difference Tests. Difference tests using χ^2 are another popular approach to measurement invariance testing. In this approach, each model with the new constraints is tested against the previous model. We did not incorporate χ^2 difference testing in our original preregistration because previous research has shown the χ^2 difference test has an inflated type I error rate when sample sizes are large (e.g., Hu & Bentler, 1999) and the sample sizes we used here are much larger than often used in measurement invariance testing. Furthermore, additional simulation studies have shown the χ^2 difference test overall inflates type I errors in measurement invariance testing (Counsell et al., 2020).

For completeness, we present the results of χ^2 difference tests below. We preregistered using robust maximum likelihood estimation to account for the Likert-type response (following Li, 2016), which requires either Satorra-Bentler scaled χ^2 difference test (Satorra & Bentler, 1994) or a likelihood ratio test. As the Satorra-Bentler test is underpowered in instances where sample sizes (ostensibly per group) are less than $N = 1{,}000$ (Li, 2016), we opted for the likelihood ratio test. The results of this test of invariance suggests that equality of factor loadings is not met, (χ^2_{LR} (4) = 23.97, $p < .001$; Table 5), and thus, the search for meaning in life/gavagai scales are not measurement equivalent.

Thus, using the χ^2 difference test method, we would not see full measurement invariance. Although the measures of model fit are all good, in such a situation one can look for partial invariance or using other methods like alignment models (which we pursue below).

Partial Invariance. One approach when full measurement invariance fails is to pursue partial measurement invariance. Partial measurement invariance is the process of estimating factor means across groups while allowing noninvariant items to either have unique factor loadings in groups, unique intercepts in a group, or both (e.g., Byrne et al., 1989). While partial measurement invariance has received criticisms (e.g., Robitzsch & Lüdtke, 2022); it is still a popular method of dealing with full measurement noninvariance.

The steps for measurement noninvariance involve first when encountering evidence that there is measurement noninvariance in either factor loadings or intercepts, to look to see if a small number of items may be causing the problem, then testing if whether imposing the more constrained model with those item(s) allowed to vary across groups can achieve measurement noninvariance.

In our confirmatory analyses, we did not find evidence for measurement noninvariance using the CFI or RMSEA, nor did we find it in the exploratory analysis (prompted by Lasker, 2022) using the TLI. This further attests to the strong evidence of measurement invariance across the search for meaning in life/gavagai scales. However, as the Likelihood Ratio tests of the chi-square values using robust maximum likelihood estimation (χ^2_{LR}) did show evidence of measurement noninvariance, we use this metric for testing partial measurement invariance.

For partial loading (metric) invariance, the largest modification index suggested the first item ("I am looking for something that makes my life feel [meaningful./gavagai]") could possibly improve model fit if the loading were allowed to freely vary between groups ($\lambda_{\mathrm{Meaning}}$ = .811, $\lambda_{\mathrm{Gavagai}}$ = .913; M.I. value = 3.845). We therefore freed the factor loading of this item and tested partial measurement invariance using the χ^2_{LR} procedure.

This approach, however, also indicated that the model with partial measurement invariance was not a better fitting model (Log-likelihood of model: −8,827.3, scaling correction factor: 2.64, # free parameters: 24) than the configural model, Log-likelihood of model: −8,817.66, scaling correction factor: 2.38, # free parameters: 30; χ^2LR (4) = 14.37, p = .006. Thus, using the chi-square likelihood ratio test, we could establish partial measurement invariance by freeing the item with the most different factor loading in the search for meaning in life ($\lambda_{\mathrm{search}}$ = .81) versus the search for gavagai scale ($\lambda_{\mathrm{gavagai}}$ = .91). Although a caveat is needed, the model had problems converging with a nonpositive definitive matrix. Instead of pursuing another item to free up the factor loading between the two scales, we instead pursue a different procedure for testing and estimating measurement invariance.

Alignment Models. Alignment models refer to a class of statistical tools that attempt to find within a given factor structure across a number of groups the right combination of variables that allow for invariant comparison across groups. This can mean that in each of the different groups, there is a somewhat unique pattern of variables retained, while still being able to compare latent means (see Luong & Flake, 2023 for an excellent tutorial). The alignment method is not typically used to confirm invariance when no noninvariance is found; rather, it is used as an alternative to the traditional MI testing approach when noninvariance is detected. The benefits of Alignment models are many, including being able to account for small amounts of noninvariance across many groups. We report the alignment method here, to further show that the rote application of statistical methods is not sufficient to conclude common measurement

Table 6. Factor Loadings (λ) Across the Five Search for Meaning in Life and Search for Gavagai Items Using the Alignment Method.

Item	$\lambda_{Meaning}$	$\lambda_{Gavagai}$	$\lambda_{Meaning}$ - $\lambda_{Gavagai}$	$SE\lambda_{\text{Meaning}}$ - $\lambda_{Gavagai}$	p	R^2
1	1.34	1.5	.153	.054	.005	< .001
2	1.53	1.48	−.05	.03	.085	.98
3	1.44	1.52	.084	.04	.033	.82
4	1.61	1.51	−.101	.03	.003	.94
5	1.62	1.58	−.046	.03	.105	.98

This method uses a α = .001 for rejecting invariance, and higher R^2 values are indicative of stronger evidence for invariance. As can be seen, the alignment method supports measurement invariance of factor loadings.

is occurring. As the alignment method tests every combination of variables, there is a large possible of type I error inflation; we therefore set the rejection threshold for noninvariance at α = .001 (following Asparouhov & Muthén, 2014).

As we confirmed that both the Search for Meaning in Life and Gavagai contained a single factor each, we used the single-factor model of each. Alignment allows for factor mean comparisons while accounting for small amounts of measurement noninvariance. We first examine the comparisons for loadings and intercepts to identify any noninvariant items.

Factor Loading Invariance. There was no evidence that factor loadings differed between the search for meaning/gavagai factors for any items at $p < .001$ (see Table 6). Furthermore, as all items except one had a high R^2, this stands as evidence for invariance, even if some items do not have high R^2 values (Muthén & Asparouhov, 2018). Therefore, using the alignment method, we are able to show that there was measurement invariance between the search for meaning in life and search for gavagai scales, which of course is ridiculous.

Item Intercept Invariance. The next step in the alignment procedure, when factor loading invariance holds (as we showed) is to test item intercept invariance. There was evidence that one of the item intercepts differed at $p < .001$ (item 5, I am searching for meaning/gavagai in my life). As this is only one item out of five, however, this does not exceed 25% of the scale's items (e.g., Luong & Flake, 2023). Whatever bias may be present with this noninvariant item will not meaningfully affect interpretation of factor means. Furthermore, the presence of high R^2 values indicates evidence for invariance in the rest of the items in the alignment model (see Table 7).

Factor Mean Comparison. Having established good evidence for measurement invariance between the search for meaning in life and search for gavagai scales using the alignment method using cutoffs suggested in the literature, we have evidence for a scale that can be meaningfully compared in our two versions—which is ridiculous. This is happening when the gavagai version of the scale is not measuring anything meaningful in the first place. Thus, traditional interpretation of measurement invariance using the alignment method suggests we can compare the latent means. Here

Table 7. Item Intercepts (τ) Across the Five Search for Meaning in Life and Search for Gavagai Items Using the Alignment Method.

Item	$\tau_{Meaning}$	$\tau_{Gavagai}$	$\tau_{Meaning}$ - $\tau_{Gavagai}$	$SE\tau_{\text{Meaning}}$ - $\tau_{Gavagai}$	p	R^2
1	5.02	4.88	−.139	.047	.003	.97
2	4.83	4.83	.003	.03	.914	1
3	4.91	4.87	−.037	.03	.253	1
4	4.76	4.81	.044	.03	.175	.99
5	4.72	4.91	.191	.042	$< .001$	-

This method uses an α = .001 for rejecting invariance, and higher R^2 values are indicative of stronger evidence for invariance. As can be seen, the alignment method supports measurement invariance of item intercepts except for item 5 (I am searching for meaning/gavagai in my life), meaning the resulting scale is within tolerable levels of invariance with only 1.5 items showing noninvariance (see Luong & Flake, 2023). People indicated they were searching for significantly more gavagai in their life than meaning in life.

we show that people are searching for gavagai less than they are searching for meaning in life ($\Delta M_{\text{Meaning-Gavagai}} = -.58, p < .001$). This is, of course, nonsensical—but that is the point. Measurement invariance achieved through the alignment method, as we have shown here, is also insufficient to make comparisons between groups.

Additional Exploratory Measures. At the suggestion of one reviewer, we also try to identify careless responding in the data that could have been brought on by the use of the gavagai manipulation. As this study was a randomized controlled trial, we look to see if seeing the gavagai items caused changes in careless responding. Following recommendations for post hoc checking (e.g., Meade & Craig, 2012; as most corrections for careless responding or response styles involve methods that need to be included before collecting data, see also Van Vaerenbergh & Thomas, 2012, for this point) we look at straightlining (answering all questions straight down) as a metric of inattentive responding (response time could not be used as time per question or page was not recorded). Using the *careless* package in *R* (Yentes & Wilhelm, 2018) on the search for meaning in life/gavagai items, we find participants were more likely to straightline in the Gavagai condition (M = 4.37, SD = 1.38, n = 601) than if they had read the proper Search for Meaning in Life questions, M = 3.22, SD = 1.51, n = 782, $t(1{,}341.2) = -14.66, p < .001, d = .79$.

Discussion

Here we show that measurement invariance can occur in the absence of any sort of meaningful measurement—which is problematic. If one wishes to interpret measurement invariance as evidence that one is *measuring the same thing*, measurement invariance cannot occur when one is clearly *not* measuring the same thing. Measurement invariance is, at most, a necessary but surely not a sufficient condition for arguing one is measuring the same thing.

This idea was first shown with the curious finding of measurement invariance on a math test given to second graders and college freshmen (Widaman et al., 1992). While the freshmen scored better and finished faster, measurement invariance was seen despite putatively different processes being used to answer the questions (e.g., math operations vs. memory). Such evidence, though interesting is not definitive, as numerous other processes could account for the findings.

Here we randomly assigned American adults to answer either a standard scale or a scale measuring nothing. Despite the large sample size and confidence that we were not measuring anything, we found measurement invariance between the factors measuring the search for meaning in life and a nonsense factor.

This finding is noteworthy because it shows strong evidence against the statement *if not measuring the same thing, then measurement invariance will not occur*; which is logically equivalent to *If you achieve measurement invariance then you are measuring the same thing* (through contraposition). Thus, our findings, and the additional evidence we reviewed, show there is strong evidence against the interpretation of measurement invariance meaning you are *measuring the same thing*. At the very least, we should refine our conception of measurement invariance away from *measuring the same thing* and instead focus more on the direct interpretation of this scale shows the same psychometric properties in these two contexts/samples. The work here shows and reinforces that measurement invariance is at best a necessary condition for equal measurement, not a sufficient condition. Meaning, just because you can establish measurement invariance does not mean you are measuring the same thing.

How Was Measurement Invariance Achieved?

In our preregistered test, we found evidence for strong measurement invariance between a standard personal beliefs scale and a nonsense scale. This was shown using the procedure for testing measurement invariance by restricting the model to have the same loadings and intercepts within a cutoff of ΔCFI and ΔRMSEA worsening no more than .01. Two exploratory tests using χ^2 difference testing and ΔTLI indices did not show evidence of strong measurement invariance, however. These two metrics were not chosen a priori because of their record of having an inflated type I error rate (e.g., Hu & Bentler, 1999; Sass et al., 2014). Whether these rejections of measurement invariance are indeed type I errors is unknowable in principle. Which set of results one chooses to believe cannot be based on its fit with a preferred outcome, however.

Either way, if strong measurement invariance was indeed initially rejected, follow-up procedures for comparing latent means could be applied though either partial measurement invariance or using the alignment method. When we applied the alignment method in a further exploratory analysis, we found evidence for measurement equivalence and the conditions to compare latent means (using cutoff criteria advanced by Asparouhov & Muthén, 2014), see also Flake & Luong, 2023). This should not be.

The problem is that in the gavagai condition, nothing is being consistently measured across participants (although participants could impute their own meaning, this was not consistent across participants, some imputed meaning, some sex, some other things, for example). And yet, all the statistical trappings of a fully functional scale have appeared, including good model fit, strong factor loadings (replicating Maul, 2017), and covariance with similar yet conceptually distinct constructs (albeit with a smaller covariance). Measurement invariance should not even come close, as measurement is not occurring.

All that is required to show evidence of adequate factor model fit is within-person consistency and between-person variance (see Maul, 2017, for this argument). So as long as participants are not answering randomly (breaking within-person consistency), and people differ in how they answer, factor models will fit data to some extent. This can be mistaken for measurement.

Response styles are one such problem, where people have different styles of responding. Some people avoid the endpoints of scales, some people stick to the middle of scales, some people avoid negatives and negative numbers (see Van Vaerenbergh & Thomas, 2013, for example). This creates problems for scale design because with the presence of within-person consistency and people using different styles, the appearance of measurement can occur. This is why in one study, when presenting completely blank items to participants, a single factor model could still be constructed with adequate model fit properties (Maul, 2017).

Our exploratory analysis suggested the gavagai condition led to more straightlining than had appeared in the search for meaning in life condition. This is what led to the trappings of measurement, between-person variance (not everyone straightlined), and within-person consistency (the response style).

Steps Forwards

One way out of this problem of measurement invariance without measurement is to increase the strictness of the criteria for showing measurement invariance. Although we used cutoffs commonly applied and shown to be robust in the literature (e.g., ΔCFI and ΔRMSEA worsening no less than .01; Cheung & Rensvold, 2002), later criteria has actually eased the statistical cutoff of ΔRMSEA to worsen no more than .015 (e.g., Chen, 2007). The evidence here would obviously pass this less-stringent criterion. We do not believe the solution is changes to cutoff levels—as cutoffs have been backed up by extensive simulation studies and are widely used. Furthermore, even if we assume a stricter requirement, a goalpost that could be moved in this one situation, our follow-up investigation using alignment modeling still was able to ‘‘show’’ a comparable scale and mean comparison between the two conditions. Instead, we first advocate abandoning the idea that measurement invariance means you are measuring the same thing, then, we take a more critical approach to measurement in general.

The limitation that led to this nonsensical measurement invariance is the rote application of statistical tests. The simple fact may be that measurement invariance testing is a necessary but not sufficient condition for claiming common measurement or comparing means between groups.

Future Directions

Previous research has called into question how to interpret noninvariance when it does occur; for example, measurement noninvariance can occur when one group is either at ceiling or floor effects, due to a necessary constraint on between-person variance in the floored and ceiling items (Welzel , Brunkert, Kruse, & Inglehart, 2023; Welzel Kruse, & Brunkert, 2023).

Overall, it seems that measurement invariance is at best a necessary condition for showing evidence that one is measuring the same thing in different populations. Showing that one has achieved measurement invariance, however, is insufficient for claiming common measurement. Whether this is a shortcoming of latent variable modeling as an accurate accounting of measurement, or simply a problem of philosophical interpretation of invariance, remains to be seen. What can be shown, however, is that one can achieve good psychometric properties, high scale reliability, and measurement invariance, in a scale that is not measuring anything at all.

One may believe that the scale to which our nonsense scale was adapted, the search for meaning in life, is likewise nonsense. If this were the case, however, we would not expect there to be a relationship between the search for meaning in life and other beliefs (such as the belief in free will). As such a relationship between search for meaning in life and the belief in free will was not only found but was particularly strong (r = .43; Gignac & Szodorai, 2016), it is unlikely the Search for Meaning in Life is a nonsense scale. That measurement invariance could be found with that scale and our actually nonsense gavagai scale, even after removing anyone who intuited the gavagai scale as being about meaning in life, remains difficult evidence against the common measurement interpretation of measurement invariance.

It is important to acknowledge that participants were imposing their own interpretation onto the sentences with the word gavagai, but in the analyses, those who interpreted the sentences to be about meaning were removed from the analysis. This helped ensure whatever was being measured in the gavagai group, if anything was being measured at all, could not be consistent across participants and was not *directly* meaning in life. It could be that there were a sort of network of constructs being tapped, including those asked about (e.g., happiness; money; sexual pleasure) and constructs not asked about. The interconnectedness of these constructs could lead to elements of ‘‘meaning’’ being present due to their associative nature, potentially allowing for measurement invariance. In short, meaning could have been *sort of* measured—*indirectly*, leading to measurement invariance. Future studies using this setup of nonsense items could explore crafting questions that keep the structure of the original scale but lack inherent associations with the original scale. In other words, the questions do not hint to the survey-taker what the blank could be in a nonsense condition.[3]

Conclusion

The process of measurement is difficult, even more so when the phenomena to be measured are unobservable mental events. Latent variable modeling is an extremely useful concept to ground our practices in; and measurement invariance testing is an important step in ensuring fair and valid test use. Yet showing measurement invariance is insufficient for claiming one is measuring the same thing. Stricter tests of the underlying assumptions of psychological measurement will help us refine and improve as a science.

Acknowledgment

The authors thank Andy Maul for his help and advice on this project and Jordan Lasker for his reanalyses and challenges.

Declaration of Conflicting Interests

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

Funding

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

ORCID iD

John Protzko https://orcid.org/0000-0001-5710-8635

Notes

[1] One need not fully stop when model fit reduces beyond a given threshold but can pursue partial measurement invariance (see Byrne et al., 1989; Jung & Yoon, 2017, for examples). This sub-step, however, is not relevant to our argument here.

[2] PsychInfo database was used instead of Google Scholar because Google Scholar's search results are tailored based on the users personal search history, so searches cannot be reproduced across user accounts. PsychInfo does not have this limitation.

[3] We thank an anonymous reviewer for this suggestion.

References

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

Borsboom, D. (2005). Measuring the mind: Conceptual issues in contemporary psychometrics. Cambridge University Press.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, *110*(2), 203–219. https://doi.org/10.1037/0033-295X.110.2.203

Bowers, E. P., Li, Y., Kiely, M. K., Brittian, A., Lerner, J. V., & Lerner, R. M. (2010). The five Cs model of positive youth development: A longitudinal analysis of confirmatory factor structure and measurement invariance. *Journal of Youth and Adolescence*, *39*(7), 720–735. https://doi.org/10.1007/s10964-010-9530-9

Buss, A. R., & Royce, J. R. (1975). Detecting cross-cultural commonalities and differences: Intergroup factor analysis. *Psychological Bulletin*, *82*(1), 128–136. https://doi.org/10.1037/h0076156

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

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 464–504.

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology*, *95*(5), 1005.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Counsell, A., Cribbie, R. A., & Flora, D. B. (2020). Evaluating equivalence testing methods for measurement invariance. *Multivariate Behavioral Research*, *55*(2), 312–328. https://doi.org/10.1080/00273171.2019.1633617

Cunningham, W. R. (1980). Age comparative factor analysis of ability variables in adulthood and old age. *Intelligence*, *4*(2), 133–149. https://doi.org/10.1016/0160-2896(80)90011-2

Dolan, C. V., Colom, R., Abad, F. J., Wicherts, J. M., Hessen, D. J., Van, De, & Sluis, S. (2006). Multi-group covariance and mean structure modeling of the relationship between the WAIS-III common factors and sex and educational attainment in Spain. *Intelligence*, *34*(2), 193–210. https://doi.org/10.1016/j.intell.2005.09.003

Dweck, C. S., Chiu, C. Y., & Hong, Y. Y. (1995). Implicit theories and their role in judgments and reactions: A word from two perspectives. *Psychological Inquiry*, *6*(4), 267–285. https://doi.org/10.1207/s15327965pli0604_1

Flake, J. K., & Luong, R. (2023). Measurement invariance testing using confirmatory factor analysis and alignment optimization: A tutorial for transparent analysis planning and reporting. *Psychological Methods*, *28*, 905–924.

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, 102, 74–78. https://doi.org/10.1016/j.paid.2016.06.069

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), Syntax and semantics, 3: Speech acts (pp. 41–58). Academic Press. https://doi.org/10.1163/9789004368811_003

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*(3), 117–144. https://doi.org/10.1080/03610739208253916

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

Irvine, S. H. (1969). Contributions of ability and attainment testing in Africa to a general theory of intellect. *Journal of Biosocial Science*, 1(S1), 91–102. https://doi.org/10.1017/S0021932000023245

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*(4), 409–426. https://doi.org/10.1007/BF02291366

Jung, E., & Yoon, M. (2017). Two-step approach to partial factorial invariance: Selecting a reference variable and identifying the source of noninvariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(1), 65–79. https://doi.org/10.1080/10705511.2016.1251845

Kim, S., Vandenabeele, W., Wright, B. E., Andersen, L. B., Cerase, F. P., Christensen, R. K., & . . .Palidauskaite, J. (2013). Investigating the structure and meaning of public service motivation across populations: Developing an international instrument and addressing issues of measurement invariance. *Journal of Public Administration Research and Theory*, *23*(1), 79–102. https://doi.org/10.1093/jopart/mus027

Lasker, J. (2022, April 20). Measurement invariance testing works. https://doi.org/10.31234/osf.io/r7e6f

Li, C. H. (2016). The performance of ML, DWLS, and ULS estimation with robust corrections in structural equation models with ordinal variables. *Psychological Methods*, *21*(3), 369–387. https://doi.org/10.1037/met0000093

Little, T. D. (2013). *Longitudinal structural equation modeling*. Guilford Press.

Luce, R. D., Suppes, P., & Krantz, D. H. (2007). Foundations of measurement: Representation, axiomatization, and invariance (Vol. *3*). Courier Corporation.

Luong, R., & Flake, J. K. (2023). Measurement invariance testing using confirmatory factor analysis and alignment optimization: A tutorial for transparent analysis planning and reporting. *Psychological Methods*, *28*(4), 905.

Markus, K. A., & Borsboom, D. (2013). Frontiers of test validity theory: Measurement, causation, and meaning. *Routledge*. https://doi.org/10.4324/9780203501207

Marsh, H. W., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A. J. S., Trautwein, U., & Nagengast, B. (2010). A new look at the Big-Five factor structure through exploratory structural equation modeling. *Psychological Assessment*, *22*, 471–491. https://doi.org/10.1037/a0019227

Marsh, H. W., Muthén, B., Asparouhov, A., 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*, 439–476. https://doi.org/10.1080/10705510903008220

Maul, A. (2017). Rethinking traditional methods of survey validation. *Measurement: Interdisciplinary Research and Perspectives*, *15*(2), 51–69. https://doi.org/10.1080/15366367.2017.1348108

McGraw, B., & Jöreskog, K. G. (1971). Factorial invariance of ability measures in groups differing in intelligence and social economic status. *British Journal of Mathematical and Statistical Psychology*, *24*(2), 154–168. https://doi.org/10.1111/j.2044-8317.1971.tb00463.x

McNeish, D., & Wolf, M. G. (2021). Dynamic fit index cutoffs for confirmatory factor analysis models. *Psychological Methods*, *28*, 61–68. https://doi.org/10.1037/met0000425

McNeish, D., & Wolf, M. G. (2022). Dynamic fit index cutoffs for one-factor models. *Behavior Research Methods*, *55*, 1157–1174. https://doi.org/10.3758/s13428-022-01847-y

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455. https://doi.org/10.1037/a0028085

Meredith, W. (1964a). Notes on factorial invariance. *Psychometrika*, *29*(2), 177–185. https://doi.org/10.1007/BF02289699

Meredith, W. (1964b). Rotation to achieve factorial invariance. *Psychometrika*, *29*(2), 187–206. https://doi.org/10.1007/BF02289700

Michell, J. (2021). Representational measurement theory: Is its number up? *Theory & Psychology*, *31*(1), 3–23. https://doi.org/10.1177/0959354320930817

Millsap, R. E. (1997). Invariance in measurement and prediction: Their relationship in the single-factor case. *Psychological Methods*, *2*(3), 248–260. https://doi.org/10.1037/1082-989X.2.3.248

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.

Nadelhoffer, T., Shepard, J., Nahmias, E., Sripada, C., & Ross, L. T. (2014). The free will inventory: Measuring beliefs about agency and responsibility. *Consciousness and cognition*, 25, 27–41.

Pages, R., Protzko, J., & Bailey, D. H. (2022). The breadth of impacts from the abecedarian project early intervention on cognitive skills. *Journal of Research on Educational Effectiveness*, *15*(2), 243–262.

Protzko, J. (2016). Does the raising IQ-raising g distinction explain the fadeout effect?. *Intelligence*, 56, 65–71.

Protzko, J., Zedelius, C. M., & Schooler, J. W. (2019). Rushing to appear virtuous: Time pressure increases socially desirable responding. *Psychological science*, *30*(11), 1584–1591.

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*, 41, 71–90.

Robitzsch, A., & Lüdtke, O. (2022). Why measurement invariance is not necessary for valid group comparisons. https://psyarxiv.com/cjyqp

Rock, D. A., Werts, C. E., & Flaugher, R. L. (1978). The use of analysis of covariance structures for comparing the psychometric properties of multiple variables across populations. *Multivariate Behavioral Research*, *13*(4), 403–418. https://doi.org/10.1207/s15327906mbr1304_3

Sass, D. A., Schmitt, T. A., & Marsh, H. W. (2014). Evaluating model fit with ordered categorical data within a measurement invariance framework: A comparison of estimators. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(2), 167–180. https://doi.org/10.1080/10705511.2014.882658

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variable analysis: Applications for developmental research* (pp. 399–419). Sage.

Steger, M. F., Frazier, P., Oishi, S., & Kaler, M. (2006). The meaning in life questionnaire: Assessing the presence of and search for meaning in life. *Journal of Counseling Psychology*, *53*(1), 80–93. https://doi.org/10.1037/0022-0167.53.1.80

Steinmetz, H., Schmidt, P., Tina-Booh, A., Wieczorek, S., & Schwartz, S. H. (2009). Testing measurement invariance using multigroup CFA: Differences between educational groups in human values measurement. *Quality & Quantity*, *43*(4), 599–616. https://doi.org/10.1007/s11135-007-9143-x

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 38, 1–10. https://doi.org/10.1007/BF02291170

Uher, J. (2021). Psychometrics is not measurement: Unraveling a fundamental misconception in quantitative psychology and the complex network of its underlying fallacies. *Journal of Theoretical and Philosophical Psychology*, *41*(1), 58–84. https://doi.org/10.1037/teo0000176

Van Vaerenbergh, Y., & Thomas, T. D. (2013). Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research*, *25*(2), 195–217. https://doi.org/10.1093/ijpor/eds021

Vernon, P. E. (1967). Working papers on cross-cultural applications of factor analysis [Mimeographed Report]. Institute of Education, University of London.

Vernon, P. E. (1969). Intelligence and cultural environment. Methuen.

Welzel, C., Brunkert, L., Kruse, S., & Inglehart, R. F. (2023). Non-invariance? An overstated problem with misconceived causes. *Sociological Methods & Research*, *52*(3), 1368–1400.

Welzel, C., Kruse, S., & Brunkert, L. (2023). Against the Mainstream: On the Limitations of Non-Invariance Diagnostics: Response to Fischer et al. and Meuleman et al. *Sociological Methods & Research*, *52*(3), 1438–1455.

Wetzel, E., Brown, A., Hill, P. L., Chung, J. M., Robins, R. W., & Roberts, B. W. (2017). The narcissism epidemic is dead; long live the narcissism epidemic. *Psychological Science*, *28*, 1833–1847. https://doi.org/10.1177/0956797617724208

Wetzel, E., Donnellan, B., & Trzesniewski, H. (2018). Generational Changes in Self-Esteem and Narcissism. *The SAGE Handbook of Personality and Individual Differences: Volume II: Origins of Personality and Individual Differences*, 132–145.

Whitehurst, G. J., Arnold, D. S., Epstein, J. N., Angell, A. L., Smith, M., & Fischel, J. E. (1994). A picture book reading intervention in day care and home for children from low-income families. *Developmental Psychology*, *30*(5), 679–689. https://doi.org/10.1037/0012-1649.30.5.679

Widaman, K. F., Ferrer, E., & Conger, R. D. (2010). Factorial invariance within longitudinal structural equation models: Measuring the same construct across time. *Child Development Perspectives*, *4*(1), 10–18. https://doi.org/10.1111/j.1750-8606.2009.00110.x

Widaman, K. F., Little, T. D., Geary, D. C., & Cormier, P. (1992). Individual differences in the development of skill in mental addition: Internal and external validation of chronometric models. *Learning and Individual Differences*, *4*(3), 167–213. https://doi.org/10.1016/1041-6080(92)90002-V

Yentes, R. D., & Wilhelm, F. (2018). careless: Procedures for computing indices of careless responding (R package version 1.2). https://github.com/ryentes/careless/.