Article

Cross-Cultural Research
2025, Vol. 59(2) 147–179
© 2025 SAGE Publications
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/10693971241312459
journals.sagepub.com/home/ccr
Sage

Why We Need to Rethink Measurement Invariance: The Role of Measurement Invariance for Cross-Cultural Research

Ronald Fischer[1,2], Johannes A. Karl[2,3], Markus Luczak-Roesch[2,4], and Larissa Hartle[1,5]

Abstract

Claims about human behavior have been hampered by limited availability of comparable data across cultures. Invariance testing has been proposed to address questions about the comparability of data, yet statistical methods have been challenged on various grounds across psychology and related fields. We highlight how current debates confuse distinct issues and fail to consider the role of data within science. We aim to overcome the impasse by a) summarizing various criticisms, b) distinguishing five mapping processes that occur during a typical research project and c) exploring how thinking about invariance as mapping can move current discussions forward. Specifically, we differentiate 1) mapping ideas to theoretical constructs and concepts, 2) mapping constructs to stimuli, 3) mapping participants' responses to stimuli

[1]Instituto D'Or for Research and Teaching, São Paulo, Brazil
[2]Victoria University of Wellington, Wellington, New Zealand
[3]Stanford University, Stanford, CA, USA
[4]Te Pūnaha Matatini, Auckland, New Zealand
[5]IDOR/Pioneer Science Initiative, Rio de Janeiro, Brazil

Received: 11 October 2024; accepted: 20 December 2024

Corresponding Author:
Ronald Fischer, Instituto D'Or for Research and Teaching, Av. Brigadeiro Luís Antônio, 5001 - Jardim Paulista, São Paulo 01401-002, Brazil.
Email: ronald.fischer@idor.org

onto numerical representations, 4) testing internal relations of stimuli responses (the typical focus of statistical invariance testing) and 5) mapping empirical observations back to theoretical statements. We treat invariance testing as a theory-guided process that offers important insights about instruments, construct validity and psychological theories throughout the research process that are currently missed by focusing on only the statistical details. Psychological claims that are valid for all humans depend on questions of invariance in the broad sense that we outline here.

Keywords

equivalence, invariance, cross-cultural comparisons, measurement, operationalization, mapping, human universals

Understanding the human mind and behavioral manifestations in different social, economic, and cultural settings presents fascinating intellectual and methodological quandaries for psychological researchers. The comparability of empirical observations, such as individuals' expressed support for specific policies, their expressed level of mental health, or endorsement of personal and cultural values demands scrutiny to ensure accurate representation of the intended theoretical construct across diverse populations. What does a score of 3.2 in culture A versus 4.1 in culture B on a scale from 1 to 5 tell us about the level of Extraversion in these two cultures? Considering these points raises fundamental questions about the conceptual nature of measures and what they capture about the psychological phenomenon of interest. What do these measures tell us about the mind, perception, motivation, or cognition of individuals around the world?

Statistical invariance testing has been the technical solution to address concerns about the comparability of data, examining whether measurement properties of observed scores are transportable or generalizable across different samples or populations (Leitgöb et al., 2023; Meredith, 1993; Vandenberg & Lance, 2000). The issue of invariance testing using complex factor analysis models nevertheless remains controversial. First, if measurement invariance is considered quintessential to research, lack of measurement invariance is considered to preclude any comparisons across groups (e.g., gender, age, ethnicity, culture, experimental conditions) or time (Boer et al., 2018; Leitgöb et al., 2023; Maassen et al., 2023; Meuleman et al., 2022). It is often claimed that this perspective has become the mainstream of psychological research (Gardiner & Funder, 2023). However, this normative claim is strongly contradicted by the observation that that the majority of currently published articles ignore measurement invariance (Boer et al., 2018; Maassen et al., 2023; Van Herk & Goldman, 2022). Furthermore, given the challenges of achieving full measurement invariance, a number of researchers have recently argued that demanding measurement invariance is counterproductive because it restricts cross-cultural research and is even irrelevant for empirical research (Funder & Gardiner, 2024; Robitzsch & Lüdtke, 2023; Welzel et al., 2021).

We aim to contribute to this ongoing debate by summarizing criticisms of the statistical invariance paradigm and responding to these points by outlining five mapping processes that take place during empirical research to ensure comparability of theory and data extending beyond the statistical aspects (elaborating on a recent introduction of these ideas by Fischer & Rudnev, 2024). We use the term 'mapping' because mathematically it refers to the process of assigning objects in one particular set to objects in another set (or possibly the same set). It is a useful term and metaphor because researchers tend to make claims about one set of objects (e.g., psychological concepts) based on information derived from another set of objects (e.g., distributions of item responses in a survey). Importantly, these mappings processes concern the equivalence across theory, operationalization, data and interpretations and not only the statistical question of internal consistency of scales. We argue that explicating the conceptual issues of transportability or generalizability of psychological constructs across populations during the research process allows for important insights about the unity versus diversity of the human mind, going beyond the idea of merely determining whether statistical invariance is achieved to enable interpretation of results (our view resembles Meehl, 1990, our distinct focus is on the applicability of theoretical constructs and concepts across cultural groups). If a personality instrument fails to demonstrate generalizability across numerous samples, it could indicate measurement problems. Alternatively, it could imply theoretically that individuals construct their personalities differently, implying a form of cultural relativity. This epistemological uncertainty by necessity requires contemplation throughout the research process to avoid erroneous conclusions. Reflecting on these processes can help researchers to critically evaluate their assumptions, refine theory and measurement approaches, and advance our knowledge of the human condition.

Basic Steps in Statistical Invariance Testing

Statistical invariance testing is typically discussed from a latent variable perspective, associated with factor analytical and item response theory models. There are three levels of invariance[1] commonly discussed in the literature (Leitgöb et al., 2023; Meredith, 1993; Vandenberg & Lance, 2000). First, if the same items can be used to measure a theoretical variable in different groups, we have configural invariance. For example, does the item "I feel blue" effectively indicate Depression in all samples studied? The direction of item loadings (the extent to which each item measures Depression) is consistent across groups, but specific loading strength or item discrimination parameters may vary between samples.

Second, if the relative loading patterns of factors (or item discrimination) are identical across samples, we have metric invariance. This means the item discriminates equally well among individuals with the same underlying trait, and it is equally related to the latent variable in all samples. Metric invariance is typically understood to allow for comparing correlations and mean patterns across cultural samples that have been included in the same invariance analysis.

Third, if intercept parameters or item difficulty are identical across samples, we have scalar invariance. This ensures that the slopes between items and latent variables are identical, and items are equally easy or difficult overall. Scalar invariance allows for direct mean comparisons and interpretation based on the assumed psychological construct.

Statistical methods differ in their ability to identify item bias and invariance as well as in their relative strictness (Bauer et al., 2020; Boer et al., 2018; Fischer & Karl, 2019; Leitgöb et al., 2023; Maassen et al., 2023). Our aim here is not to discuss these various approaches as this is well covered in other sources. Just as a note to frame our following discussion, a broad Confirmatory Factor Analysis (CFA) approach has become (for better or worse) the de-facto gold standard for invariance testing (Leitgöb et al., 2023). Regardless the method, the major challenge is to decide what to do when invariance tests identify problems with either a set of parameters or some samples.

A Brief Overview of Criticisms of Invariance Testing

Several practical, conceptual or statistical issues of invariance testing have been identified in the literature which we will outline in the following section.

Difficulties in Establishing Invariance with Increasing Number of Samples

A practical concern is that multigroup CFA applied to datasets with more than a few samples is inevitably leading to the rejection of measurement invariance (Funder & Gardiner, 2024; Welzel et al., 2021; for a more positive assessment, see Van Herk & Goldman, 2022). Establishing scalar invariance in this context is particularly challenging. The implication of violating scalar invariance is that cross-cultural comparison of scale scores is open to alternative interpretations (Fontaine, 2005; Leitgöb et al., 2023; Vandenberg & Lance, 2000).

Lack of Clarity Around Standards and Criteria

There is a lack of clarity around measurement standards and statistical criteria (Funder & Gardiner, 2024). Because there a) is no agreement among experts regarding testing guidelines, and b) simulation studies suggest complex design-issues and variable thresholds for making decisions on model parameters (Hu & Bentler, 1999; Kang et al., 2016; Marsh et al., 2004; Rutkowski & Svetina, 2014), the complexity may discourage researchers to apply these techniques.

Invariance as Binary Choice

Current practice treats invariance as a binary of accepting or rejecting invariance (Fischer & Karl, 2023; Nye & Drasgow, 2011). Therefore the identification of non-invariance often deals the death blow to empirical research (Fischer & Karl, 2023; Gardiner & Funder, 2023; Karl & Fischer, 2022), foreclosing possibly interesting avenues for cultural analyses. Hence, invariance testing is seen as a barrier for research (Welzel et al., 2021).

Invariance Testing Could Increase Cultural Bias in Instruments

A central issue in any attempt to measure concepts across cultures is the profound cultural bias in contemporary psychology (Boehnke, 2022). Western instruments are typically translated into different languages and applied in regions of the world where the content of the measurement instrument may not speak to the concerns, priorities, or realities of the research participants. During test development or adaptation, non-invariant items are typically excluded from the instrument. Boehnke (2022) challenged that this leads to culturally biased instruments, because content is removed from the instrument that may be relevant in at least one setting, but not the others. The remaining content may not speak to specific concerns of the non-Western, non-WEIRD population because typically little effort is made to develop items that are locally meaningful. This further contributes to cultural bias and domain underrepresentation of the theoretical construct space (Fontaine, 2005).

Statistical Criteria do not Address Semantic (Content) Sameness

Invariance is treated as a statistical question which assumes that the quantitative results carry the same semantic meaning, yet this semantic sameness is not explicitly tested by the statistical test (Boehnke, 2022; Meehl, 1978; Robitzsch & Lüdtke, 2023). Translation checks are often assumed sufficient to imply that the semantic content is adequately covered. However, the statistical analysis only proves that the observed covariance metrics are equivalent. As famously noted by Lord: “The numbers don’t remember where they came from” (Lord, 1953, p. 751).

Invariance Testing Being Irrelevant for Examining Culture-Level Dynamics

A further conceptual criticism is that instruments intended to measure culture-level processes do not require individual-level invariance testing because the level of analysis is different and therefore, invariance tests are irrelevant (Welzel et al., 2021; Welzel & Inglehart, 2016). In this line of reasoning, collective level processes should not be evaluated at the individual level, even if the indices that make up collective level constructs are measured at the individual level. Validity is supposedly guaranteed by demonstrating meaningful correlations with external variables at the aggregate level (for a related point, see Funder & Gardiner, 2024).

Restriction of Variance

Invariance testing relies on examinations of variability. An argument against invariance testing when examining culture-level processes is that variability in group means measured on Likert-type scales with limited response options (typically 5, 7 or 10-point scales) leads to variance restrictions if group means move to either end of the response scale due to cultural processes. Hence, cultural dynamics may shift observed means to either end of a scale which then reduces available variance and causes bias in invariance tests conducted at the individual level (Welzel et al., 2021).

Invariance as a Latent Variable Problem

Invariance testing is typically discussed within a latent variable paradigm, which assumes that there is an underlying psychological variable that 'causes' the behavioral observations captured by each item or stimuli (most commonly associated with factor analysis). As a result, other conceptualization of theoretical constructs, in particular emergent, formative or second-order structures in which independent processes converge and lead to the emergence to a new construct (e.g., socio-economic status is a classic example) are thought to be outside the scope of invariance testing (Welzel et al., 2021; Welzel & Inglehart, 2016).

Items are not Interchangeable

One basic assumption within a multigroup CFA approach is that the items should be interchangeable, that is they are assumed to be randomly selected from a universe of items and locally independent, controlling for the implied latent variable. However, most instruments have a limited number of items which are a) unlikely to represent a random selection out of a universe of indicators and b) not conditionally independent of each other given the latent variable (Robitzsch &

Lüdtke, 2023). Considering the limited content covered by a small number of items within most measures, removing one or more items will likely changes the meaning of the construct being measured (Singh, 1995; Steenkamp & Baumgartner, 1998; Van Herk & Goldman, 2022) and the interpretation of any pairwise difference becomes open to alternative interpretations. Robitzsch and Lüdtke (2023) argue that the interpretation of a misspecified (that is poorly fitting) multigroup factor model is conceptually cleaner than a well-fitting model in each sample that relies on different parameters in each sample.

Circularity in the Absence of External Validation Standards

Invariance testing involves circularity to some extent. A latent variable itself cannot be independently verified in the statistical model and must always be inferred from manifest external variables. Therefore, the interpretation of non-invariance becomes circular as the latent variable cannot be examined absent independent information (Robitzsch & Lüdtke, 2023).

Invariance Testing Ignores Causal Structures

The causal modeling literature has pointed out that statistical criteria used for invariance tests are unable to differentiate between conflicting causal claims (VanderWeele, 2022; VanderWeele & Vansteelandt, 2022). Even when there is evidence of an invariant measurement model across cultures, it does not carry any information about the causal structures within and across cultures (Sterner et al., 2024). The common assumption is that underlying latent variables rather than the individual indicators are causally relevant. The construction of a single score subsequent to invariance testing leaves open the equally plausible option that causal effects are mediated via the indicators rather than the latent variable.

Summary of Criticisms

There is an increasing uneasiness about the demands of invariance tests, in particular the rigid application of the multigroup CFA model. In our view, each of these claims have some merit but they tend to miss the larger point about the issue of comparability of theory and observation in cross-cultural research. Some of the points have also been raised in relation to psychological measurement in general (Alexandrova & Haybron, 2016; Maraun, 1998; Meehl, 1990), so our points could be considered as a crucial extension for developing psychology as a science relevant for all of humanity.

In light of the poor reporting of invariance testing in the current literature, these increasingly vocal challenges to invariance testing take on new urgency. In the following section, we take up the valid criticisms and show how they can be reinterpreted by considering research as a series of mapping processes.

Some Important Definitions in Our Argument

We need to briefly define some key terms for our ongoing discussion, to help us link theory to empirical observations. We call a theoretical idea that researchers develop to facilitate understanding of the world and how the world operates a ***construct*** (Lambert & Newman, 2022; Podsakoff et al., 2016). The Latin root of the term (*con* together and *struere* to build or pile) nicely captures both the agentic component as well as the joining of diverse thought patterns into a more complex theoretical idea. A second term that is often used interchangeable with construct is "concept" (Latin, *concipere* to conceive). We use the term ***concept*** to refer to a more singular idea, that is more directly linked to some observable unit in the real world. Podsakoff et al. (2016) defined concepts as 'cognitive symbols (or abstract terms) that specify the features, attributes, or characteristics of the phenomenon in the real or phenomenological world that they are meant to represent and that distinguish them from other related phenomena' (p. 161). Concepts in our terminology can be seen as forming part of a larger construct, such as the concepts of talkativeness, impulsiveness, sociability, and dominance all forming part of Extraversion as a construct (see Figure 1).

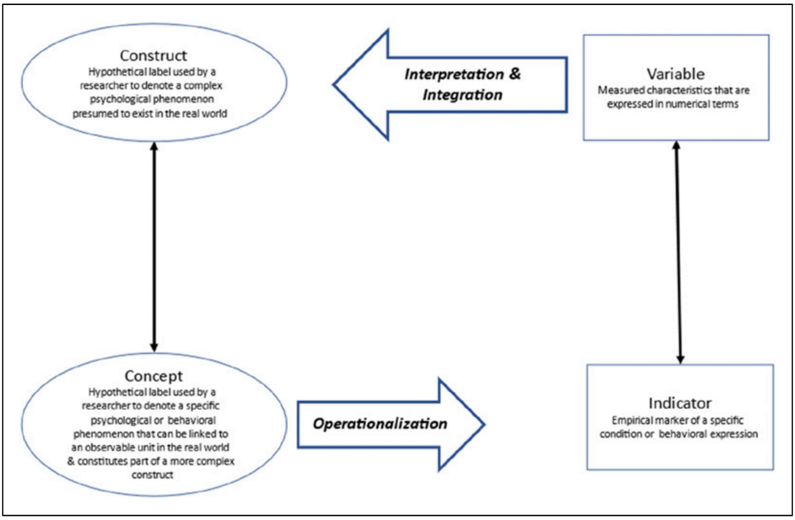
Construct
Hypothetical label used by a researcher to denote a complex psychological phenomenon presumed to exist in the real world
Interpretation & Integration
Variable
Measured characteristics that are expressed in numerical terms
Concept
Hypothetical label used by a researcher to denote a specific psychological or behavioral phenomenon that can be linked to an observable unit in the real world & constitutes part of a more complex construct
Operationalization
Indicator
Empirical marker of a specific condition or behavioral expression

Figure 1. Visualization of the relationship between as-intended constructs and concepts and their relationship with empirical variables (measured constructs) and their indicators, items or stimuli (measured concepts).

Both constructs and concepts can be differentiated into an as-intended and as-determined type within the operationalization process (Haucke et al., 2021). The as-intended refers to the construct or concept as the idea(s) to be investigated in a study. In contrast, the as-determined type refers to the actual measurement that is taking place during the study and serves as the operational definition of the as-intended. The as-determined construct is often considered a variable, as it represents the measured characteristics of a person or object that are expressed in different numerical terms (e.g., categories or levels). The as-determined concept can be thought of as an indicator, stimuli or item, which is the empirical marker to elicit the concept as-intended.

This distinction can be traced to broader debates in the philosophy of science. Meehl (1990) observed that substantive theories involve various auxiliary theories in order to state empirical observations or make empirical predictions, some of which are intertwined with measurement procedures. This amounts to a differentiation between core theoretical statements and auxiliary or peripheral elements of a theory (cf. Lakatos, 1970). We could argue that the clear specification of core aspects of the theoretical construct and how they may be instantiated via concepts and indicators in different cultural contexts is essential for theoretical claims to progress. It is important to clearly state what is core about a theoretical process and allow for modification of peripheral elements to improve explanatory power (e.g., the concept of 'truth-likeness' or verisimilitude). The exploration of what may need to be adapted at the peripheral level may even be the starting point for a stronger theoretical account (see below the fifth mapping step).

Five Major Mapping Processes in Culture-comparative Research

In the following, we outline five distinct mapping processes (see Figure 2 and Table 1) that researchers routinely engage in when doing culture-comparative research. We explain how explicit attention to each of these processes could address criticisms related to invariance, while acknowledging some of the shortcomings of the current statistical invariance paradigm.

First Mapping – Ideas to Constructs

How can we make sense of the thoughts, beliefs, and emotions that others are experiencing? As researchers we need to consider whether our constructs are meaningful and relevant in different cultural contexts. We argue that a significant number of criticisms of the invariance approach could be resolved if explicit definitions were established during this initial mapping and subsequent invariance testing decisions were guided accordingly. We elaborate on

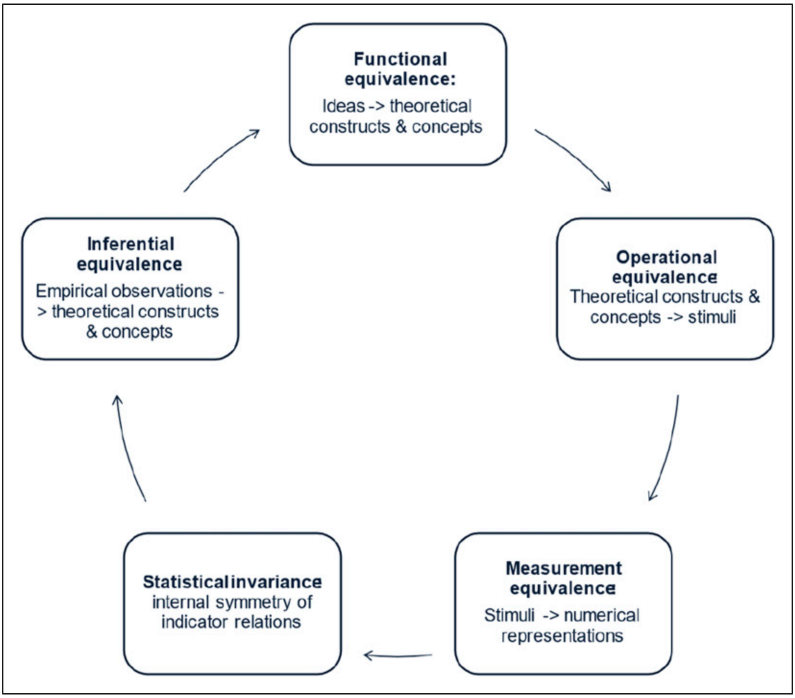
Functional equivalence:
Ideas -> theoretical constructs & concepts
Operational equivalence
Theoretical constructs & concepts -> stimuli
Measurement equivalence
Stimuli -> numerical representations
Statisticalinvariance
internal symmetry of indicator relations
Inferential equivalence
Empirical observations -> theoretical constructs & concepts

Figure 2. Schematic representation of the mapping steps. Note: This is a schematic representation, and we do not imply unidirectionality of the individual steps: research often is iterative and cyclical with insights gained at a later stage requiring a reconsideration of earlier steps (for example insights from a pilot test may result in renewed theorizing and subsequent changes in operationalization).

this first mapping because it is often neglected, yet it is the foundation of all subsequent mappings.

Functional Equivalence of Constructs

The question of functional equivalence (Fontaine, 2005) examines whether the same concepts and constructs are relevant across-cultures for accounting for empirical observations of the theoretical construct (Fontaine, 2005).

To highlight one specific challenge in culture-comparative research is that linguistic terms that denote a specific theoretical concept in one language may not be directly available in another. For example, there has been much recent discussion about emotions and cultural relativism (Jackson et al., 2019; Sauter et al., 2015). It may be easy to claim cultural relativism when a specific term is not available in one language. However, the idea of functional equivalence

Table 1. An Overview of Major Mapping Processes in Culture-comparative Research.

Mapping	Technical term	Definition	Relevant criticism of invariance approach	Steps to address criticisms
Ideas - > theoretical constructs and concepts	Functional equivalence	Relative equality of theoretical constructs and concepts embedded within a theoretical nomological network	Cultural bias, level of analysis, ontological nature of constructs, causal structure, circularity, binary choice as barrier	Define constructs & concepts within their cultural context Clarify the ontological nature of constructs (common cause, common effect, reciprocal, etc.) Define appropriate levels involved in the process (e.g., individual, institutional, ecological, societal) Consider domain representation, breadth and relevance Consider linguistic, social, economic, technological, etc. conditions
Theoretical constructs and concepts - > stimuli	Operational equivalence	Relative equality of the extent to which theoretical concepts are translated into empirical stimuli that evoke the relevant and representative information about the intended theoretical concept	Cultural bias, semantic content, interchangeability of items	Specify linguistic, social, economic, technological, etc. conditions that may modify how constructs and concepts are expressed

(continued)

Table 1. (continued)

Mapping	Technical term	Definition	Relevant criticism of invariance approach	Steps to address criticisms
Stimuli - > numerical representations	Measurement equivalence	Relative equality of symbolic (numerical) representations of a theoretical concept across different measurement occasions	Restrictions of variance, binary choice, number of samples, lack of clarity	Explore variability in numeric representations & verbal labels for gradients, intensities, frequencies across cultures, scaling properties may not be aligned with the functionality of the psychological construct/ concept
Internal relations among indicators	Statistical invariance testing	Extent of symmetry of a set of observations (measurements) of concepts in relation to each other across measurement occasions	Ontological nature of constructs, interchangeability of items, restriction of variance, binary choices as barrier, lack of clarity	Standard statistical considerations of the data, including underlying assumptions of test, alignment of the assumptions underlying the statistical test with ontological nature of the construct/concept
Empirical observations - > theoretical constructs	Inferential equivalence	Holistic consideration and integration of available empirical information across all samples and an evaluation of the implications for functional equivalence	All criticisms	Explain what the data patterns imply about the theoretical concept, what theoretical arguments given the operationalization can be conditionally accepted and what theoretical arguments need to be refuted?

shifts the focus to the degree of emotional similarity in behavioral responses to situational stimuli, which may allow researchers to identify a common core that is independent of linguistic idiosyncrasies. It requires careful consideration of the evidence, but it is possible to find that individuals in regions of the world that are highly culturally distinct experience comparable behavioral reactions to scenarios, even if the specific emotion word is not available to all groups (Breugelmans & Poortinga, 2006). It is at this level that criticisms of cultural bias need to be first addressed at a conceptual level. If, after conceptual elaboration, statistical invariance is still rejected in a research project, the focus may shift to developing theoretical arguments as to why some constructs may function differently in some cultural contexts compared to others, and what factors may be relevant in explaining this difference.

The Ontological Nature of Constructs - Linking Concepts and Constructs

A common assumption in psychology is the hierarchical nature of constructs: concepts (e.g., talkativeness) form part of more complex constructs (e.g., Extraversion[2]). How do individual concepts relate to each other and the overall construct? There are at least three different ontological interpretations to explain variations between observations in empirical data in the social sciences (Kruis & Maris, 2016).

The first and most common in psychology is a reflective, latent variable or common cause model. An unobserved latent variable that causes the observed states in a measured concept is proposed, therefore, observations are seen as indicators of the underlying causal process (Borsboom et al., 2003).

A second interpretation is a formative, common effect, combinatory logic model or collider model. Here, independent observations in their combination give rise (or collide) into an emerging effect (Edwards & Bagozzi, 2000).

Finally, a reciprocal effect or fully connected conditional network model presumes no ‘unobservables’ or larger constructs as such and instead associations between observations are seen as the consequence of reciprocal or mutualistic associations between the observations (Borsboom et al., 2021).

Given certain assumptions, the three interpretations are mathematically equivalent and substantive reasoning needs to be used to determine their logical plausibility (Kruis & Maris, 2016; Marsman et al., 2018). As a result, empirical observations may be compatible with any of the implied interpretations, and this forces researchers to specify the theoretical model up front. For example, is Extraversion a latent variable that causes individuals to talk more (e.g., talkativeness)? Or do various independent concepts such as talkativeness and being outgoing when meeting strangers create the emergence of a new concept of Extraversion in those individuals where these traits tend to co-occur (Cramer et al., 2012)? Or could we potentially think of

Extraversion as a cluster of tightly coupled concepts such as talkativeness, impulsivity and being outgoing, which in their overall clustering show characteristics of a behavioral syndrome? This is an important issue to carefully map out during the initial stages of a research project when selecting, refining and elaborating the theoretical background. As discussed above, the ontological nature of theoretical constructs has been used in criticisms against invariance testing, so this is an important issue to clarify upfront and to then make the relevant methodological choices in line with theory. This is at the core of recent exchanges around invariance that tap into the ontological questions of concepts for culture-comparative research (Meuleman et al., 2022; Sokolov, 2018; Welzel et al., 2021). These issues are certainly of interest and challenging for psychology more generally (Maraun, 1998; Michell, 1997). Our concern here is focused specifically on the consideration of the comparability of the theoretical constructs and concepts across cultural groups.

Domain Representation and Relevance of Concepts

A further issue within this first mapping stage when discussing cultural work is the issue of the domain representation and domain relevance of concepts (Fontaine, 2005; van de Vijver & Leung, 2021). These questions are of relevance for addressing criticisms around cultural bias as well as the disconnection between statistical criteria and semantic content. In some cultural contexts, the theoretical constructs tend to be rather broad and encompassing. Returning to our example of talkativeness, if we assume that talkativeness is an important concept within the construct of Extraversion, we may need to think about what this concept implies in different cultural contexts. What does it mean to be talkative? The amount of talking overall? The loudness? The content and focus of conversations (e.g., gossip)? The choice of words (e.g., eloquence vs swearing)? The extent of turn taking and given others a chance to speak? The context (e.g., talking in class vs at a party)? Drawing upon cultural stereotypes of talkability (for a behavioral exploration and critical view, see Stivers et al., 2009), we may ask what are possible baselines and behavioral distributions that we have to consider. What does it mean to be talkative in Nordic countries which are supposed to be rather non-expressive and sullen versus South American countries, where it seems to be common practice to interrupt and talk on top of each other?

The issue of representativeness is about whether the specific behavioral expression is seen as reflecting an important element of the concept in each of the cultural contexts. For example, how important is talkativeness for the larger construct of Extraversion? This is not a trivial question as recent network analyses of Extraversion facets across instruments in monocultural samples have suggested (Schwaba et al., 2020). Across cultural contexts, these issues are even more complex considering possible differences in communication styles, which may give different weight or even meaning to levels of talkativeness, embedded with cultural notions of power, status and hierarchy (e.g., talkativeness of a high status vs low status individual, situational awareness) (Wan, 2021).

Some concepts may be irrelevant or associated with other constructs in some contexts compared to others. For example, an exploration of personality structures in Thailand (Fischer, 2021a) suggested that talking to others can be interpreted as being dishonest or careless, leading to a different association in this highly collectivistic and tight context. Obviously, this was an empirical observation, but it demonstrates that these issues may need careful attention during the theoretical stage and lead to different psychological theories (Gurven et al., 2013). As noted by Meehl more than 45 years ago (Meehl, 1978), the theoretical elaboration of psychological constructs is typically poor and a casual reading of the contemporary research in leading psychology journals suggests that not much has changed since then (Alexandrova & Haybron, 2016).

Therefore, the criticism of disconnection between statistical criteria and semantic content and issues of cultural bias arises at a theoretical level and need to be considered from the outset of research.

Identifying the Levels of a Theory

One of the more complex issues from a cultural perspective within this first mapping stage is the specification of the appropriate level of theory – is this construct applicable to individuals, to groups or to nations (Fischer, 2009; Klein et al., 1999; Van De Vijver et al., 2008)? This is closely linked to the long-standing and non-trivial issue of the level of analysis in culture-comparative research (Hofstede, 1980), which has also created confusion and criticism for invariance testing. Culture is typically defined as a shared meaning-system, implying relative agreement around specific values, beliefs or norms within groups and differentiation across groups (Faulkner et al., 2006). However, large survey studies have demonstrated substantial individual variability and relatively weak cultural differences, strongly undermining the case for consensus and a shared meaning system (Fischer, 2021b; Fischer & Schwartz, 2011; Saucier et al., 2015).

To cite one prominent example, Schwartz (2014) proposed a latent definition of culture, in which culture is a hypothetical, latent variable situated outside the minds of individuals, but is mediated via social institutions that structure the beliefs, values and thought process of members of a society. In other words, cultural values are latent societal level constructs that causally influence institutions, which in turn influence the responses of individuals to value items in surveys, but survey responses also reflect other variables.

Nevertheless, survey responses in their aggregate are thought to capture the extent to which culture as a latent variable causally impacts all individuals within a society. These aggregated 'means themselves are not the cultural values, but they are observable consequences from which we infer cultural values' (p. 5).

One criticism is that it invokes the famous ghost in the machine (Morris, 2014), invoking an improbable device to solve a seemingly unsolvable problem. Morris (2014) highlighted that this conceptualization of culture becomes a near caricature in its unidirectional top-down operation of cultural processes and that this approach fails to account for cultural change. If we assume that 'no individual carries the culture. The culture influences every individual in a unique way' (Schwartz, 2014, p. 5), then taken to its logical extreme, culture becomes irrelevant for psychologists interested in individual differences.

What is the point of us discussing this literature here? First and foremost, it requires a principled theoretical argument about the construct that a researcher is trying to capture conceptually when including individuals from different populations. What a researcher's conceptual position on the nature of the theoretical construct? Does it rely on individual-level psychological processes that explain variability both within and across human populations? Or is it a characteristic of institutions that have no psychological equivalent at the level of the individual, but nevertheless influence or constrain individual behavior? If the latter, how does the theoretical account explain the variability in individual responses, if such responses are being used to infer characteristics of the institution? If researchers are interested in using this definition of culture and general approach, a clear explanation of theoretical mechanisms linking the different levels is needed (for a related argument, see Sterner et al., 2024). It is not convincing to argue that invariance tests are inapplicable because of mismatched levels and to ignore discussions of the causal structures that generate the data.

Explicating Causal Relationships

This last point needs broader attention: Any theoretical argument about the nature of reality requires some statement about the wider implied causal web of relationships to understand plausible antecedents and effects of the construct of interest (Deffner et al., 2022; Sterner et al., 2024). Which variable is an antecedent and therefore is expected to lead to causal changes in other variables? As these are theoretical arguments, there needs to be a clear causal line of reasoning to avoid ambiguity (VanderWeele, 2022). This discussion is related to questions of external validity that have been recently raised in arguments against invariance testing. Clear specification also needs to address the circularity criticism: Observations need to be tied to a causal theory. We agree that questions about the underlying causal structure are an important theoretical problem (Alexandrova & Haybron, 2016; Moreau & Wiebels, 2022) and this requires careful attention at the outset of culture-comparative research .

Second Mapping – Concepts to Stimuli and Indicators

Concepts in our definition are theoretical statements about ideas that have a concrete expression in the real world. As an empirical science, the theoretical ideas in psychology need to be mapped onto stimuli that elicit a reaction that can indicate the measurable level of the theoretically implied concept. This is the classic process of operationalization (Haucke et al., 2021). Attention to operationalization is crucial when examining concepts across populations. Here we use the term operational equivalence to indicate the *relative equality of the extent to which theoretical concepts are translated into empirical stimuli that evoke the relevant and representative information about the intended theoretical concept*. These issues are central for addressing criticisms around cultural bias, semantic content and the interchangeability of items.

Let us consider a few of the indicators that are often included in Extraversion-type tests (see the Supplement for item examples). As part of the first mapping, researchers needed to think about the theoretical as-intended concept. What does talkativeness mean in relation to behavior (frequency, amplitude, or number of interaction partners)? The focus in widely used surveys for culture-comparative research is quite narrowly focused on frequency (talkative, talking a lot) and an enjoyment aspect (like or enjoy talking). If we think about talkative as an adjective, it may refer to an inability to be concise (chattering or wordy), which carries a certain negative evaluative connotation (e.g., being chatty, big-mouthed, glib), but could also include a certain erudite connotation (e.g., being articulate, fluent, or eloquent). The same term may therefore invoke differential evaluative qualities which may or may not be intended by the researchers. Such qualitative nuances in evaluation are exaggerated crossing cultural boundaries and when translating the term into a second language. For example, in Portuguese talkative could be *falador/ faladeira* or even *tagarela*, which are increasingly negative terms. In Samoan the most direct translation already invokes an evaluative statement, e.g., *tautalatala* – too much talk. In German, it could be a rather positive *gesprächig* or a rather negative *geschwätzig*. In some German dialects such as Bavarian or Saxon, an appropriate translation may be *gschnadderig* or *schnatterisch*, which is rather negative and invokes absent-minded animal-like behavior ("geese chatter)". A single word may take on additional meaning when translated, as the brief examples of translations of the term talkative have shown.

Providing behavioral or situational anchors may help but involves additional trade-offs. Contextual elements may help distract from the intended content and make indicators more vulnerable to extraneous cultural differences. For example, the NEO item implies an emphasis on the pleasurable aspect (really enjoy), which includes the element of volition and positive reinforcement experienced by engaging in the activity of talking. The Eysenck measure specifies both the pleasurable aspect as well as the target – namely strangers. The item that forms part of the compassion aspect of Agreeableness instead focuses on the content of the conversations, that is the wellbeing of others. It is not necessarily clear that these various indicators capture the same theoretical idea, even though they are using conceptually related words.

By specifying the exact content and context (e.g., talking at parties, talking to strangers), indicators may also invoke additional norms or schemas. For example, talking to strangers can have different connotations depending on one's age, gender, and social status in different parts of the world. In some traditional societies in the Pacific, the village chief is the person who must receive strangers first to ascertain the intentions, thus limiting the options for interactions for most individuals. In many cultures around the world, there are explicit norms and restrictions for cross-gender interactions. When referring to specific interaction environments (e.g., parties), this also then invokes possibly different connotations of what it means to have a party. These issues tap into but certainly go beyond the widely discussed guidelines for translation (Chidlow et al., 2014; Harkness et al., 2003).

It is also increasingly common to use secondary data which raises interesting conceptual questions because the theories of the original researcher may not align with the theories of researcher re-using that data. In many cases, it may be more economical or simply more practical to use existing data, than to embark on a costly data collection from scratch that is more clearly aligned with the theoretical as-intended constructs. In these cases, the researchers must redouble their efforts to a) explicate their theoretical models, b) do their due diligence in understanding the data and its operationalization process and c) link the empirical indicators back to plausible theoretical interpretations in terms of constructs and concepts.

The point of the second mapping is to think through the behavioral instantiations of the theoretical concept, attempting to translate the as-intended idea into an as-determined behavioral stimuli, be it for developing research designs from scratch or be it mapping theoretical ideas onto pre-existing indicators and data sets. Again, question of cultural bias in the development and adaptation of measures together with questions about the semantic content and the interchangeability of items need careful considerations that go beyond but do not preclude tests of statistical invariance.

Third Mapping – Stimuli to Numerical Representations

The third mapping deals with the mapping of stimulus responses to numeric scores utilized by researchers. Inspired by earlier work that had highlighted the implications of this measurement process for cross-cultural comparisons (Van De Vijver & Poortinga, 1982) we define measurement equivalence as *the relative equality of symbolic (numerical) representations of a theoretical concept across different measurement occasions*. In other words, it is concerned with the uncertainties that a researcher has about interpreting numerical representations of theoretical concepts implemented in a specific single indicator but obtained in different measurement contexts.

For example, when considering the item 'is talkative', respondents are given the task to respond to the question stem 'I am someone who…' and then mark their responses on a scale from 1 'disagree strongly' to 5 'agree strongly'. Intermediate responses are scored as 2 'disagree a little', 3 'neutral, no opinion' and 4 'agree a little'. This measurement scale is used to map a theoretical concept such as talkativeness onto a numerical scale. Therefore, the third mapping question is whether this item adequately captures the relative position of individuals.

The question of equivalence of numerical scores is not tied to a specific statistical model but rather can be traced back to ideas of scaling (Stevens, 1946), which is concerned with 'assignment of numerals to objects or events according to rules' (p. 677). Stevens popularized the now commonplace distinctions of nominal (categorical), ordinal (ranked), interval (equal distances), and ratio (same origin or zero point) scales.

Let us briefly consider the implications for comparative measurement. Within such a comparative context, the concern is whether the same numerals can be assigned to concepts following the same rules. The level of functional equivalence most clearly maps onto nominal scaling properties – the concept can be captured or not. At the simplest level talkativeness can either take one of two classes – yes (it exists in culture X) or no (it does not exist in culture X).

Next, structural equivalence addresses whether the same observed variables can be used to measure the intended concept as discussed above (the second mapping). Under structural equivalence, we could assume that an ordinal relationship exists between numerical scores on a specific measure and the intended theoretical concept, expressing a rank-ordered relationship which orders the observations in the same way according to the theoretical concept across cultural groups. For example, the question 'is talkative' can be used to order individuals in the same way across groups according to their theoretical level of talkativeness. No claims beyond relative ordering of participants can be made.

Metric equivalence conceptually corresponds to interval scales. Under metric equivalence, numerical scores across groups preserve equality of intervals or differences in the theoretical construct. With this assumption, a difference between a score of 4 and 5 can now be interpreted meaningfully across two or more groups. However, it is not yet possible to make any claims about the meaningfulness of a score of 4 on a 1 to 5 scale across groups.

To make claims about the means, it is necessary to assume scalar equivalence, which corresponds to a ratio scale. Here, we assume the presence of an absolute zero value that is comparable across groups (although the empirical observation of the absolute zero may not be necessary). Only at this level can we make claims that two individuals from two different samples with the empirically observed score of 4 have the same level of talkativeness.

Let us emphasize a really important point here - although we have been talking about numerical scores, we have not invoked statistical techniques such as MGCFA or IRT models in this discussion. The discussion of equivalence in our definition and mapping is about what a single observation can tell us about a concept that we are trying to measure with an item. Criticisms of variance restriction or questions about the ontological status of the theoretical concept are irrelevant here. What is of interest is the relative certainty we can have about the measurement quality of each individual observation in a data set.

There are possible ways to examine the quality of this mapping process.. First, it may be possible to examine the test-retest consistency (e.g., reliability) to examine whether the ordering of individuals remains stable. This requires some assumptions about the temporal or contextual stability of the theoretical construct. Second, independent observational data could be used to validate the item responses. With the advancement of capturing digital footprints or real-world behavior via wearable sensors, triangulation of item responses against behavioral responses becomes feasible. A third option is to use qualitative methods, such as response probing or response process evaluation (Meitinger & Behr, 2016; Taves et al., 2021). All of these methods are important at the theoretical level, and should be explored in their own right and/or to complement, as a follow-up or to interpret invariance testing results.

Fourth Mapping – Symmetry of Internal Relations

The fourth mapping involves an exploration of the relationship between observed scores (measured concepts) in relation to their presumed shared construct. Returning to our example item ‘is talkative’, an instrument could include three other items that are thought to capture the same construct of ‘sociability’. For simplicity, let us consider just one additional indicator here: ‘is outgoing, sociable’. By examining the relative performance of these two indicators in relation to each other (or in relation to all other items within the trait facet), we can gain some insights into the relative validity of the measurement process. As the focus is now on the performance of items, this is often called differential item functioning when considered in the context of multiple measurement occasions (Zumbo, 2007). What is important is the relative functioning of the items in relation to all other items.

The statistical process of examining these properties is invariance testing. We define invariance as the *extent of symmetry of a set of observations (measurements) of concepts in relation to each other across measurement occasions*. Do measured indicators show similar patterns relative to each other in different cultural contexts? The most common approach are latent variable statistical tests, but other models can equally address questions of internal symmetry which overcomes one of the criticisms of invariance testing (Tantardini et al., 2019; van Borkulo et al., 2022). Decisions on structural, metric, and scalar invariance are based on statistical properties derived from the intercorrelations of the individual items (or intercepts) with the overall score. Therefore, the focus shifts towards the identification of differential item functioning relative to each other and tests only reveal whether a set of items shows internal relations that differ across groups. As noted in the criticisms outlined above, judging the internal structure of instruments is quite a different question compared to what we want to know, namely, whether the measurement process itself is free of bias.

Responding to binary choice criticisms, invariance as symmetry of relations among indicators is a question of degree and should not be treated as a binary categorial decision of invariant or not invariant. Focusing on the extent to which parameters vary across populations and across indicators is theoretically more informative and insightful rather than focusing a binary decision that may preclude further inquiry (Fischer & Karl, 2023; Karl & Fischer, 2022; Nye & Drasgow, 2011). There is an increasing number of options to both examine effect sizes of non-invariance (Nye & Drasgow, 2011) as well as options for modeling the non-invariance of both individual items and overall non-invariance parameters (Bauer et al., 2020; Davidov et al., 2012; Fischer & Karl, 2023; Fontaine, 2008; Zyphur et al., 2008). This is an important avenue for linking the statistical analysis of invariance parameters back to theoretical questions that can advance our understanding of cultural processes and overcome the perception that non-invariance is a mechanical step and a barrier for meaningful research.

Statistical options exist for latent variable (Boer et al., 2018; Fischer & Karl, 2019; Leitgöb et al., 2023; Vandenberg & Lance, 2000), formative (Adamantios & Papadopoulos, 2010; Henseler et al., 2016) and network models (van Borkulo et al., 2022), overcoming a core criticism of invariance as only being relevant or available for latent variable models. Furthermore, options for non-parametric invariance testing that overcome variance restrictions are available (Bauer et al., 2020).

However, these statistical approaches are complex and more simulation research needs to examine cut-off criteria for larger number of samples. The complexity of state-of-the-art approaches is a fair criticism of the literature and more work for both establishing and communicating best practice as well as clear criteria needs to be done.

Fifth Mapping – Empirical Observations Back to Theory

The fifth and final mapping requires mapping empirical observations back onto the theoretical constructs and concepts that a researcher set out to explore. This is closely linked to functional equivalence but working in reverse via a holistic consideration and integration of available empirical data across all samples and an evaluation of the theoretical implications. Statistical information from the measurement invariance test is informative, but the focus should be on the theoretical implications and compatibility with pre-existing information. A failure to identify a certain level of invariance (e.g., lack of internal symmetry of factor loadings or intercepts) across a larger number of cultural samples is an empirical observation that needs theoretical consideration. For this reason, we call this interpretative equivalence because it concerns the quality of interpretations possible in relation to theory and observations across the different cultural contexts.

The interpretation of statistical invariance information within the context of the theoretical framework may vary. For example, statistical non-invariance in a formative (common effect) versus reflective (common cause) model has different implications. In a common effect model, non-symmetry in emergence of a theoretical construct across different cultural contexts implies different patterns in the emergence of a theoretically relevant variable. Different behaviors in one context converge around the interpretation of a behavior cluster as indicating an extraverted individual, whereas in a different context a slightly different combination of behaviors is interpreted as a socially engaged (vs disengaged) individual (that is an individual who is interested in the wellbeing of others, which involves additional elements of both Agreeableness and Openness in addition to Extraversion, leading to a qualitatively different personality trait (Fischer, 2021a; Gurven et al., 2013)).

In a reflective or common cause model, non-invariance may indicate different cognitive processes are used by respondents in relation to the indicators or stimuli, which may imply that the same problem is solved through different psychological mechanisms in different populations or that there are situational constraints on the expression of the common factor. For example, if talkative does not load together with other Extraversion items, a reflective model interpretation could be that there are situational constraints on the behavior in this context that do not allow for an expression of the latent variable in this particular behavior (Gurven et al., 2013). This of course would be highly relevant for a cultural analysis and imply rather different cultural processes or dynamics.

To put it explicitly, the same statistical phenomenon of non-invariance has different theoretical implications when considered from a formative versus reflective model. In both cases, there is a theoretical implication that needs to be considered, e.g., personality emerges differently due to distinct constellations and combinations of behavioral traits versus situational variables that modify or suppress the expression of a presumed latent behavioral predisposition. Of course, there may also be clear problems with the operationalization of key variables, i.e. problems with the methodological implementation of the theoretical ideas. Classic issues are translation problems or errors in the application of the instruments across cultural contexts. A careful integration of both theoretical arguments and empirical evidence is needed to make claims in one direction or another.

In our view, this integrative mapping stage has to move beyond typical calls for external validity (Funder & Gardiner, 2024; Welzel et al., 2021; Welzel & Inglehart, 2016). The relevant mapping issue is the reconnection of empirical data to theoretical constructs and linking observed correlations to causal claims made within a theoretical space. If the issue of external or predictive validity (e.g., predicting some other variable of interest in real-world context) is central to the theory, then this needs to be incorporated into the research process from the start (Fischer & Rudnev, 2024). If validity concerns are used as an afterthought in light of perceived unfavorable results from a measurement invariance analysis, then the shift towards a different criterion for evaluating the data is not going to resolve the theoretical issues implied by the invariance test result. To reiterate a key point, invariance tests provide useful information for the researcher, but their implications depend on the theoretical perspective that needs to be spelled out in the initial mappings of the research process.

This integrative mapping also needs to consider the origin of the data, especially when using secondary data sources for theoretical projects that involve different as-intended constructs. An important consideration are alternative explanations of the observed patterns, careful analyses of possible biases and thorough quality checks on the data. Observing significant correlations is often seen as a foregone conclusion on the quality of the data, but in a social world where nearly everything correlates with everything, a significant correlation may actually be rather meaningless (Meehl, 1978). Recent work in the philosophy of science has pointed out that poorly defined variables often correlate more consistently or more strongly with all sorts of other variables, due to their poor conceptualization and inclusion of non-specific variance (Alexandrova & Haybron, 2016). Similar, validating existing measures blindly against other measures without a careful theoretical rationale is likely to lead to a local optimum that ultimately fails to accurately describe the phenomenon of interest (Moreau & Wiebels, 2022). Lord's classic statement that numbers do not know where they come from, which is often levelled as an argument against invariance testing, equally implies that significant associations do not indicate that the theoretical process is valid.

Invariance testing in terms of symmetry of relations will provide us with some indications about the internal structure of a set of variables, the interpretation of which may take quite different forms depending on the formal model adopted by researchers. As we had briefly outlined above, our arguments are somewhat aligned with previous criticisms of psychology (Meehl, 1990), with our concern being the applicability of theoretical claims across cultural groups. Differentiating between a core theoretical claim and peripheral or auxiliary components of the theory allows for theories to progress. To the extent that a theoretical claim clearly specifies what is central, the invariance tests can be used as a protective belt in a Lakatosian sense if the claims only concern peripheral elements that can be adjusted. As noted by Meehl, such 'strategic retreats' in the face of empirical disconfirmation can help in improving theoretical claims. We would hasten to add that this information itself can provide novel insights for the development of potentially stronger theories of culture and human behavior. To put it very bluntly, ignoring this information available to researchers is bad science.

Moving Beyond the Criticism

We summarized several criticisms leveled at invariance testing at the outset of our manuscript. What can we conclude now in light of the five-stage mapping process that we have outlined? Considering the first challenge of observing non-invariance in many cross-national studies, these findings must be considered from a theoretical perspective. If we believe that CFA is an appropriate tool for psychological research, non-invariance indicates problems with the theoretical models and/or the empirical instruments. As we have exemplified with personality terms, the operationalization process is considerably more complex and important nuances may be taken into consideration to effectively measure distinct psychological constructs or concepts in different contexts. Statistical non-invariance is worth exploring further and should not be the endpoint for research. Available tools allow researchers to explore whether there is systematic variability in these invariance patterns, which can lead to new and better theorizing. We urge a more careful exploration of these patterns.

The second issue is the uncritical application of latent variable models, which may or may not be applicable to the data and theoretical framework of the researcher. It is important to understand the assumptions of these tests and carefully consider whether these tests are appropriate for the collected data and purposes. Many alternative methods are available, and researchers have plenty of options to explore the quality of their data beyond the current fashion of using some variation of the confirmatory factor model. Of course, all of these models are ultimately linked to some kind of variability estimate (typically correlations). It does not make sense to throw out the baby with the bathwater – there are (many) problems with correlation coefficients, yet researchers continue to use correlations because it is a useful tool.

Finally, all the major issues tend to return to the issue of what tests measure – we need better theoretical discussions of our constructs, concepts and theoretical ideas to sharpen our measures (Alexandrova & Haybron, 2016). Such endeavours require a greater reconsideration of the rich insights provided by qualitative descriptions of cultural contexts, as provided by anthropologists, ethnographers, sociologists, historians but also biologists and ethologists. These sources can enrichen and strengthen our theoretical lenses, helping to provide different perspectives on the phenomenon that we are interested in understanding. This is arguably our most important point; a truly universal psychology applicable and relevant for humans across cultural contexts needs strong theory about sources and patterns of variability in human behavior, irrespective or even despite of diverse methodologies and data sources.

Author Contribution

Conceptualization: RF; Validation: JAK, MLR, LH; Visualization: RF, MLR, JAK, LH; Original draft: RF; Review and editing: JAK, MLR, LH.

Author Notes

An earlier version that prepared the ground for our arguments is available online (https://psyarxiv.com/fst9k/), the current version extends on these arguments and puts them in a broader perspective by including both the general criticisms and shortcomings of current statistical discussions of invariance and extending the idea of the research process as a mapping exercise. This is the third revised and shortened version of a longer preprint (please see the version 2 of this preprint with more examples & notes: https://osf.io/preprints/psyarxiv/3f9ca). Our arguments were shaped by and iteratively crystalized during exchanges and debates with Chris Welzel, Bart Meuleman, Misho Minkov, Boris Sokolov, Plamen Akaliyski, Ella Daniel, Anat Bardi, Maksim Rudnev, Johnny Fontaine and other participants at two different symposia held at the International Association for Cross-Cultural Psychology, events organized by the Higher School of Economics, and a research presentation at the Pontifical Catholic University of Rio de Janeiro - PUC-Rio. We are grateful for the open and constructive exchange around diverse viewpoints with our colleagues and are looking forward to collectively improving scientific methods in the social and behavioral sciences.

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 iDs

Ronald Fischer https://orcid.org/0000-0002-3055-3955
Johannes A. Karl https://orcid.org/0000-0001-5166-0728
Markus Luczak-Roesch https://orcid.org/0000-0003-4610-7244

Supplemental Material

Supplemental material for this article is available online.

Notes

1. Equivalence is the term that is preferred in the cross-cultural psychology literature (Fischer & Poortinga, 2018; Fontaine, 2005; van de Vijver & Leung, 1997), whereas invariance testing is the term preferred in the psychometric and test construction literature (Leitgöb et al., 2023; Vandenberg & Lance, 2000). It is important to highlight that the term equivalence is conceptually broader as it is not directly related to any statistical process, but rather incorporates both philosophical and conceptual questions about the construct as well as the empirical analysis of the variables (Van De Vijver & Poortinga, 1982). In contrast, invariance is directly tied to the statistical manipulation of observed scores and their properties. In this sense, invariance takes a rather narrow perspective and ignores the larger conceptual issue that is of actual interest to researchers. The question of comparability is important across the whole research process, but by using the term invariance narrowly within the context of statistical inference testing, it has shifted the discussion towards technicalities that obscure what researchers try to accomplish with their studies.
2. We use Personality as an example here because it is easily understandable, personality has been discussed from different ontological perspectives and it is likely to have a substantive cultural component. However, our arguments are relevant across all measured constructs in psychology.

References

Adamantios, D., & Papadopoulos, N. (2010). Assessing the cross-national invariance of formative measures: Guidelines for international business researchers. *Journal of International Business Studies*, *41*(2), 360–370.

Alexandrova, A., & Haybron, D. M. (2016). Is construct validation valid? *Philosophy of Science*, *83*(5), 1098–1109. https://doi.org/10.1086/687941

Bauer, D. J., Belzak, W. C. M., & Cole, V. (2020). Simplifying the assessment of measurement invariance over multiple background variables: Using regularized moderated nonlinear factor analysis to detect differential item functioning.

Structural Equation Modeling: A Multidisciplinary Journal, *27*(1), 43–55. https://doi.org/10.1080/10705511.2019.1642754

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

Boer, D., Hanke, K., & He, J. (2018). On detecting systematic measurement error in cross-cultural research: A Review and critical reflection on equivalence and invariance tests. *Journal of Cross-Cultural Psychology*, *49*(5), 713–734. https://doi.org/10.1177/0022022117749042

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

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

Breugelmans, S. M., & Poortinga, Y. H. (2006). Emotion without a word: Shame and guilt among Rarámuri Indians and rural Javanese. *Journal of Personality and Social Psychology*, *91*(6), 1111–1122. https://doi.org/10.1037/0022-3514.91.6.1111

Chidlow, A., Plakoyiannaki, E., & Welch, C. (2014). Translation in cross-language international business research: Beyond equivalence. *Journal of International Business Studies*, *45*(5), 562–582. https://doi.org/10.1057/jibs.2013.67

Cramer, A. O. J., Van Der Sluis, S., Noordhof, A., Wichers, M., Geschwind, N., Aggen, S. H., Kendler, K. S., & Borsboom, D. (2012). Dimensions of normal personality as networks in search of equilibrium: You can't like parties if you don't like people. *European Journal of Personality*, *26*(4), 414–431. https://doi.org/10.1002/per.1866

Davidov, E., Dülmer, H., Schlüter, E., Schmidt, P., & Meuleman, B. (2012). Using a multilevel structural equation modeling approach to explain cross-cultural measurement noninvariance. *Journal of Cross-Cultural Psychology*, *43*(4), 558–575. https://doi.org/10.1177/0022022112438397

Deffner, D., Rohrer, J. M., & McElreath, R. (2022). A causal framework for cross-cultural generalizability. *Advances in Methods and Practices in Psychological Science*, *5*(3), 25152459221106366. https://doi.org/10.1177/25152459221106366

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods*, *5*(2), 155–174. https://doi.org/10.1037/1082-989x.5.2.155

Faulkner, S. L., Baldwin, J. R., Lindsley, S. L., & Hecht, M. L. (2006). Layers of meaning: An analysis of definitions of culture. In *Redefining culture: Perspectives across the disciplines* (pp. 27–52). Lawrence Erlbaum Associates Publishers.

Fischer, R. (2009). Where is culture in cross cultural research? An outline of a multilevel research process for measuring culture as a shared meaning system. *International Journal of Cross Cultural Management*, *9*(1), 25–49.

Fischer, R. (2021a). Alternative four-factor structure of the Mini-IPIP in Thailand. *International Journal of Personality Psychology*, *7*, 35–42. https://doi.org/10.21827/ijpp.7.37978

Fischer, R. (2021b). Origins of values differences: A two-level analysis of economic, climatic and parasite stress explanations in the value domain. *Cross-Cultural Research*, *55*(5), 438–473. https://doi.org/10.1177/10693971211031476

Fischer, R., & Karl, J. A. (2019). A primer to (cross-cultural) multi-group invariance testing possibilities in R. *Frontiers in Psychology*, *1507*.

Fischer, R., & Karl, J. A. (2023). Niche diversity effects on personality measurement – evidence from large national samples during the COVID-19 pandemic. *Current Research in Ecological and Social Psychology*, *4*, 100116. https://doi.org/10.1016/j.cresp.2023.100116

Fischer, R., & Poortinga, Y. H. (2018). Addressing methodological challenges in culture-comparative research. *Journal of Cross-Cultural Psychology*, *49*(5), 691–712.

Fischer, R., & Rudnev, M. (2024). From MIsgivings to MIse-en-scène: The role of invariance in personality science. *European Journal of Personality*, Advanced Online Publication. https://doi.org/10.1177/08902070241283081

Fischer, R., & Schwartz, S. (2011). Whence differences in value priorities? Individual, cultural, or artifactual sources. *Journal of Cross-Cultural Psychology*, *42*(7), 1127–1144.

Fontaine, J. R. J. (2005). Equivalence. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement* (pp. 803–813). Elsevier. https://doi.org/10.1016/B0-12-369398-5/00116-X

Fontaine, J. R. J. (2008). Traditional and multilevel approaches in cross-cultural research: An integration of methodological frameworks. In *Multilevel analysis of individuals and cultures* (pp. 65–92). Taylor & Francis Group/Lawrence Erlbaum Associates.

Funder, D. C., & Gardiner, G. (2024). MIsgivings about measurement invariance. *European Journal of Personality*, Advanced Online Publication. https://doi.org/10.1177/08902070241228338

Gardiner, G., & Funder, D. (2023). MIsgivings about measurement invariance. PsyArXiv. https://doi.org/10.31234/osf.io/97cxg

Gurven, M., von Rueden, C., Massenkoff, M., Kaplan, H., & Lero Vie, M. (2013). How universal is the big five? Testing the five-factor model of personality variation among forager-farmers in the Bolivian amazon. *Journal of Personality and Social Psychology*, *104*(2), 354–370. https://doi.org/10.1037/a0030841

Harkness, J. A., Vijver, F. J. R. van de, & Mohler, P. P. (Eds.), (2003). *Cross-cultural survey methods*. J. Wiley.

Haucke, M., Hoekstra, R., & van Ravenzwaaij, D. (2021). When numbers fail: Do researchers agree on operationalization of published research? *Royal Society Open Science*, *8*(9), 191354. https://doi.org/10.1098/rsos.191354

Henseler, J., Ringle, C. M., & Sarstedt, M. (2016). Testing measurement invariance of composites using partial least squares. *International Marketing Review*, *33*(3), 405–431. https://doi.org/10.1108/IMR-09-2014-0304

Hofstede, G. (1980). *Culture's consequences: International differences in work-related values*. Sage Publications.

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

Jackson, J. C., Watts, J., Henry, T. R., List, J.-M., Forkel, R., Mucha, P. J., Greenhill, S. J., Gray, R. D., & Lindquist, K. A. (2019). Emotion semantics show both cultural variation and universal structure. *Science*, *366*(6472), 1517–1522. https://doi.org/10.1126/science.aaw8160

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

Karl, J., & Fischer, R. (2022). More than yes and No: Predicting the magnitude of non-invariance between countries from systematic features. In M. Klicperova-Baker & W. Friedlmeier (Eds.), *Xenophobia vs. Patriotism: Where is my Home?Proceedings from the 25th Congress of the International Association for Cross-Cultural Psychology*, 300, pages 1–21. International Association for Cross-Cultural Psychology. https://doi.org/10.4087/ELED5219

Klein, K. J., Tosi, H., & Cannella, A. A. (1999). Introduction to special topic forum: Multilevel theory building: Benefits, barriers, and new developments. *Academy of Management Review*, *24*(2), 243–248.

Kruis, J., & Maris, G. (2016). Three representations of the Ising model. *Scientific Reports*, *6*(1). Article 1. https://doi.org/10.1038/srep34175

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In A. Musgrave & I. Lakatos (Eds.), *Criticism and the growth of knowledge: Proceedings of the international colloquium in the philosophy of science*, 1965 (Vol. *4*, pp. 91–196). Cambridge University Press. https://doi.org/10.1017/CBO9781139171434.009

Lambert, L. S., & Newman, D. A. (2022). Construct development and validation in three practical steps: Recommendations for reviewers. *Organizational Research Methods*, 109442812211153. https://doi.org/10.1177/10944281221115374

Leitgöb, H., Seddig, D., Asparouhov, T., Behr, D., Davidov, E., De Roover, K., Jak, S., Meitinger, K., Menold, N., Muthén, B., Rudnev, M., Schmidt, P., & van de Schoot, R. (2023). Measurement invariance in the social sciences: Historical development, methodological challenges, state of the art, and future perspectives.

Social Science Research, *110*, 102805. https://doi.org/10.1016/j.ssresearch.2022.102805

Lord, F. M. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement*, *13*(4), 517–549. https://doi.org/10.1177/001316445301300401

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*. https://doi.org/10.1037/met0000624

Maraun, M. D. (1998). Measurement as a normative practice: Implications of wittgenstein's philosophy for measurement in psychology. *Theory & Psychology*, *8*(4), 435–461. https://doi.org/10.1177/0959354398084001

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing hu and bentler's (1999) findings. *Structural Equation Modeling: A Multidisciplinary Journal*, *11*(3). Article 3. https://doi.org/10.1207/s15328007sem1103_2

Marsman, M., Borsboom, D., Kruis, J., Epskamp, S., Bork, R. V, Waldorp, L. J., Maas, H. L. J. V D, & Maris, G. (2018). An introduction to network psychometrics: Relating ising network models to item response theory models. *Multivariate Behavioral Research*, *53*(1), 15–35. https://doi.org/10.1080/00273171.2017.1379379

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*(4), 806–834. https://doi.org/10.1037/0022-006X.46.4.806

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry*, *1*(2), 108–141. https://doi.org/10.1207/s15327965pli0102_1

Meitinger, K., & Behr, D. (2016). Comparing cognitive interviewing and online probing: Do they find similar results? *Field Methods*, *28*(4), 363–380. https://doi.org/10.1177/1525822X15625866

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*, 00491241221091755. https://doi.org/10.1177/00491241221091755

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, *88*(3), 355–383. https://doi.org/10.1111/j.2044-8295.1997.tb02641.x

Moreau, D., & Wiebels, K. (2022). Psychological constructs as local optima. *Nature Reviews Psychology*, *1*(4). Article 4. https://doi.org/10.1038/s44159-022-00042-2

Morris, M. W. (2014). Values as the essence of culture: Foundation or fallacy? *Journal of Cross-Cultural Psychology*, *45*(1), 14–24. https://doi.org/10.1177/0022022113513400

Nye, C. D., & Drasgow, F. (2011). Effect size indices for analyses of measurement equivalence: Understanding the practical importance of differences between groups. *Journal of Applied Psychology*, *96*(5), 966–980. https://doi.org/10.1037/a0022955

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2016). Recommendations for creating better concept definitions in the organizational, behavioral, and social sciences. *Organizational Research Methods*, *19*(2), 159–203. https://doi.org/10.1177/1094428115624965

Robitzsch, A., & Lüdtke, O. (2023). Why full, partial, or approximate measurement invariance are not a prerequisite for meaningful and valid group comparisons. *Structural Equation Modeling: A Multidisciplinary Journal*, 1–12. https://doi.org/10.1080/10705511.2023.2191292

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

Saucier, G., Kenner, J., Iurino, K., Bou Malham, P., Chen, Z., Thalmayer, A. G., Kemmelmeier, M., Tov, W., Boutti, R., Metaferia, H., Çankaya, B., Mastor, K. A., Hsu, K.-Y., Wu, R., Maniruzzaman, M., Rugira, J., Tsaousis, I., Sosnyuk, O., Regmi Adhikary, J., & Altschul, C. (2015). Cross-cultural differences in a global "survey of world views.". *Journal of Cross-Cultural Psychology*, *46*(1), 53–70. https://doi.org/10.1177/0022022114551791

Sauter, D. A., Eisner, F., Ekman, P., & Scott, S. K. (2015). Emotional vocalizations are recognized across cultures regardless of the valence of distractors. *Psychological Science*, *26*(3), 354–356. https://doi.org/10.1177/0956797614560771

Schwaba, T., Rhemtulla, M., Hopwood, C. J., & Bleidorn, W. (2020). A facet atlas: Visualizing networks that describe the blends, cores, and peripheries of personality structure. *PLoS One*, *15*(7), e0236893. https://doi.org/10.1371/journal.pone.0236893

Schwartz, S. H. (2014). Rethinking the concept and measurement of societal culture in light of empirical findings. *Journal of Cross-Cultural Psychology*, *45*, 5–13. https://doi.org/10.1177/0022022113490830

Singh, J. (1995). Measurement issues in cross-national research. *Journal of International Business Studies*, *26*(3), 597–619. https://doi.org/10.1057/palgrave.jibs.8490188

Sokolov, B. (2018). The index of emancipative values: Measurement model misspecifications. *American Political Science Review*, *112*(2), 395–408. https://doi.org/10.1017/S0003055417000624

Steenkamp, J.-B. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*(1), 78–90. https://doi.org/10.1086/209528

Sterner, P., Pargent, F., Deffner, D., & Goretzko, D. (2024). A causal framework for the comparability of latent variables. *Structural Equation Modeling: A Multidisciplinary Journal*, *0*(0), 1–12. https://doi.org/10.1080/10705511.2024.2339396

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677–680. https://doi.org/10.1126/science.103.2684.677

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., de Ruiter, J. P., Yoon, K.-E., & Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, *106*(26), 10587–10592. https://doi.org/10.1073/pnas.0903616106

Tantardini, M., Ieva, F., Tajoli, L., & Piccardi, C. (2019). Comparing methods for comparing networks. *Scientific Reports*, *9*(1). Article 1. https://doi.org/10.1038/s41598-019-53708-y

Taves, M. G. W., Elliott, I., & Maul, A. A (2021). Survey item validation. In *The routledge handbook of research methods in the study of religion* (2nd ed.). Routledge.

van Borkulo, C. D., van Bork, R., Boschloo, L., Kossakowski, J. J., Tio, P., Schoevers, R. A., Borsboom, D., & Waldorp, L. J. (2022). Comparing network structures on three aspects: A permutation test. *Psychological Methods*. https://doi.org/10.1037/met0000476

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

VanderWeele, T. J. (2022). Constructed measures and causal inference: Towards a new model of measurement for psychosocial constructs. *Epidemiology*, *33*(1), 141. https://doi.org/10.1097/EDE.0000000000001434

VanderWeele, T. J., & Vansteelandt, S. (2022). A statistical test to reject the structural interpretation of a latent factor model. *Journal of the Royal Statistical Society - Series B: Statistical Methodology*, *84*(5), 2032–2054. https://doi.org/10.1111/rssb.12555

van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research* (pp. xiii–186). Sage Publications, Inc.

van de Vijver, F. J. R., & Leung, K. (2021). V. H. Fetvadjiev, J. He, & J. R. J. Fontaine (Eds.), *Methods and data analysis for cross-cultural research* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/9781107415188

Van De Vijver, F. J. R., & Poortinga, Y. H. (1982). Cross-cultural generalization and universality. *Journal of Cross-Cultural Psychology*, *13*(4), 387–408. https://doi.org/10.1177/0022002182013004001

Van De Vijver, F. J. R., van Hemert, D. A., & Poortinga, Y. H. (Eds.), (2008). *Multilevel analysis of individuals and cultures*. Lawrence Erlbaum Associates.

Van Herk, H., & Goldman, S. P. K. (2022). The advancement of measurement invariance testing in cross-cultural research in the period 1999–2020. Executing rather than scrutinizing? In H. Baumgartner & B. Weijters (Eds.), *Measurement in* *marketing* (*19*, pp. 95–119). Emerald Publishing Limited. https://doi.org/10.1108/S1548-643520220000019005

Wan, Y. (2021). Why are they so quiet? Exploring reticent and passive east Asian ESL Students in the U.S. Classrooms. *Open Journal of Modern Linguistics*, *11*(6). Article 6. https://doi.org/10.4236/ojml.2021.116073

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

Welzel, C., & Inglehart, R. F. (2016). Misconceptions of measurement equivalence: Time for a paradigm shift. *Comparative Political Studies*, *49*(8), 1068–1094. https://doi.org/10.1177/0010414016628275

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, *4*(2), 223–233. https://doi.org/10.1080/15434300701375832

Zyphur, M. J., Kaplan, S. A., & Christian, M. S. (2008). Assumptions of cross-level measurement and structural invariance in the analysis of multilevel data: Problems and solutions. *Group Dynamics: Theory, Research, and Practice*, *12*(2), 127–140. https://doi.org/10.1037/1089-2699.12.2.127

Author Biographies

Ronald Fischer works at the Cognitive Neuroscience and Neuroinformatics Unit at D’Or Institute for Research & Teaching, Brazil. He is a Fellow of the Royal Society NZ Te Aparangi and the Association for Psychological Science. His research focuses on cultural and evolutionary dynamics of values, beliefs, personality and wellbeing.

Johannes A. Karl is a visiting researcher at the Stanford Graduate School of Business, where he studies the intersection of evolutionary psychology and cultural dynamics. His research focuses on understanding the role of mindfulness, values, and wellbeing in diverse populations and how these factors evolve and influence human behavior.

Markus Luczak-Roesch is a professor for Informatics and holder of the Char in Complexity Science at Victoria University of Wellington. He is also a co-Director at Te Pūnaha Matatini, New Zealand’s national centre of research excellence in Complex Systems. His research focuses on understanding evolutionary emergence and information dynamics.

Larissa Hartle works as a postdoctoral researcher at the Cognitive Neuroscience and Neuroinformatics Unit of the D’Or Institute for Research and Teaching in Brazil. Her research interests include well-being, belief systems, and non-ordinary experiences.