Received: 20 April 2022 | Revised: 18 December 2022 | Accepted: 17 February 2023

DOI: 10.1111/spc3.12740

ORIGINAL ARTICLE

WILEY

What's wrong with rating scales? Psychology's replication and confidence crisis cannot be solved without transparency in data generation

Jana Uher[1,2]

[1]School of Human Sciences, University of Greenwich, London, UK

[2]London School of Economics and Political Science, London, UK

Correspondence

Jana Uher, University of Greenwich, Old Royal Naval College, Park Row, London SE10 9LS, UK.
Email: mail@janauher.com

Funding information

European Commission, Grant/Award Number: EC Grant Agreement number 629430

Abstract

Quantitative explorations of behaviour, psyche and society are common in psychology. This requires methods that justify the attribution of results to the measurands (the entities to be measured, e.g., in individuals) and that make the results' quantitative meaning publicly interpretable (e.g., for decision making). Do rating scales—psychology's primary methods to generate numerical data—meet these criteria? This article summarises selected epistemological and methodological problems of rating scales that arise, amongst others, from the intricacies of language-based methods and from psychologists' challenges to distinguish their study phenomena from their means of exploring these phenomena. Failure to make this logical distinction entails that disparate scientific activities are conflated, thereby distorting scientific concepts and procedures. Rating scales promote such conflations because they serve both as description of the empirical study system (e.g., behaviours) and as symbolic study system (e.g., data variables), leaving the interpretation of each system and the mapping relations between them to raters' intuitive decisions. Verbal scales, however, have broad semantic fields of meanings, which are context-sensitive and therefore interpreted differently, and which cannot logically match the quantitative meaning commonly ascribed to the numerical scores derived from them. The ease of using verbal descriptions as means of exploration drew psychologists' attention to the conceptual-interpretive level, away from their actual study phenomena. This also led them to overlook key elements of data generation and measurement. The pragmatic necessity to analyse rating scores through between-individual comparisons entailed the erroneous assumption that psychometrics and sample-level statistics could enable measurement. Improving data analyses, as currently discussed, is therefore insufficient for overcoming psychology's crises of replication, confidence, validity and generalizability. Data generation methods are necessary that make the entire process—from the empirical study phenomena up to the results—fully transparent and traceable. This rigorous analysis of rating scales highlights important steps for future directions.

KEYWORDS
generalizability, measurement, psychometrics, replicability, transparency, traceability, validation

1 | INTRODUCTION

1.1 | Psychology: A discipline in crisis and its popular method of investigation

Rating 'scales' have changed psychology. Since their introduction (Likert, 1932; Thurstone, 1928), rating 'scales' have been hailed as instruments enabling psychological 'measurement', making psychology's study phenomena amenable to mathematical analysis. This allowed psychologists to replace controversial introspective methods with natural-science methods (e.g., hypothesis testing). The efficiency of creating large numerical data sets with ratings—nowadays conveniently administered online, reaching millions of respondents remotely (e.g., Amazon's Mechanical Turk; Anderson et al., 2018)—enabled major developments in statistical analysis.

Curiously, however, rating 'scales' in themselves remained largely unchanged, while the last century saw previously unthinkable advancements in physical measurement (e.g., distance measurement using satellites). Still today, everyday descriptions of phenomena of interest (items) are presented to respondents (raters) for judgement using predefined multi-stage answer categories (rating), commonly considered a 'scale' (indicating, e.g., levels of agreement). Meanwhile, ideas that colloquial statements or questions presented with a visualised 'scale' (e.g., five stars) could enable quantitative investigation of experience and behaviour (e.g., customer satisfaction) proliferate also outside of academia (e.g., in business) as do commercial survey platforms (e.g., Qualtrics). Ratings have become ubiquitous in everyday life. Have psychologists lost their scientific authority over rating 'scales'? What actually is these methods' scientific foundation?

Psychologists seek to establish rating 'scales' as a scientific method by means of psychometric modelling—statistical analyses that scrutinise rating data for their utility to discriminate well and consistently between cases (reliability) and in ways considered important (validity). To establish reliable and valid data sets and to analyse their empirical structures, statistics have become essential. Accordingly, current debates about problematic research practices (e.g., *p*-hacking, HARKing) and psychology's crises in replication, confidence, validation and generalisability

(Andrade, 2021; Earp & Trafimow, 2015; Yarkoni, 2022) are focussed on data analysis (Uher, 2021d, 2022a, 2022b)—as are proposals for tackling them (e.g., pre-registration, robust statistics; Nosek et al., 2015; Zwaan et al., 2017).

But psychology's problems go deeper. Large survey panels yielded problematic findings with popular 'personality' 'scales'. For example, rather than showing empirical interrelations, as required for psychometric 'scales', ratings on Big Five items targeted at the *same* 'personality' construct (e.g., 'gets nervous easily' and [inversed keyed] 'relaxed, handles stress well' for 'neuroticism') varied unsystematically, averaging zero across 25 countries. Instead of showing meaningful congruence, factor structures differed substantially between student and general public samples, between different age groups and between different countries. These findings challenge these 'scales'' reliability and validity both within and across Western and non-Western countries (Condon et al., 2021; Hanel & Vione, 2016; Laajaj et al., 2019; Ludeke & Larsen, 2017). So, what do rating 'scales' actually capture?

Applications of *quantitative methods* in psychology *in themselves* are increasingly questioned, such as regarding their underlying epistemologies and measurement theories (Barrett, 2018; Buntins et al., 2016; Michell, 1999; Tafreshi et al., 2016; Trendler, 2009, 2013; Uher, 2021c, 2022b; Westerman, 2014) or regarding the inadequacy of sample-level statistics for individual-level explorations (Lamiell, 2019; Molenaar & Campbell, 2009; Richters, 2021). Finally, psychology's neglected and insufficiently developed philosophical and theoretical foundations are identified as the root cause of its persistent crises (Danziger, 1985; Haig & Borsboom, 2008; Smedslund, 2016; Szollosi et al., 2020; Teo, 2018; Toomela, 2018; Uher, 2021b; Valsiner, 2019).

This also concerns the philosophical and theoretical foundations of research methods, especially of those used to *generate* data before these can be *analysed* (Uher, 2019, 2021a, 2022b; Valsiner, 2017; Wagoner & Valsiner, 2005). Indeed, open and 'meta-science' initiatives focus on transparency in data analyses, such as in response coding and transformation, construct operationalisation and validity, as well as in the statistical tests, coefficients and parameters used (Flake & Fried, 2020; Hardwicke et al., 2022). But what about transparency in the ways in which the *raw data in themselves* are being generated in the first place?

Psychologists specify for their rating 'scales' item wordings, answer formats (e.g., five-point agreement 'scale'), instructions for administration and scoring as well as psychometric properties of rating data thus-obtained. But rating 'scales' in themselves cannot produce any data. This is done by respondents! Instructions to the persons being asked to complete rating 'scales' (raters), however, are often surprisingly vague (e.g., 'there is no right answer'). Indeed, colloquial wordings of rating 'scales' are regarded as sufficiently self-explanatory for enabling laypeople to generate data for scientific studies. But how do raters actually understand and use such 'scales'? What do they consider in their ratings? How do they reach their overall quantitative judgements and decide which answer boxes to tick? Despite the long-standing use of ratings and the identification of countless rater biases and various mental processes involved (Podsakoff et al., 2003; Tourangeau et al., 2000), a general theory about quantitative data generation with rating 'scales' for the purposes of measurement has not yet been developed (caution: not to be confused with psychometric modelling, see below; Uher, 2021c, 2021d).

1.2 | The present critical analyses and their conceptual foundation

This article presents an introductory overview of selected key problems in the epistemological and methodological foundations of rating 'scales' that have been explored using the Transdisciplinary Philosophy-of-Science Paradigm for Research on Individuals (TPS-Paradigm[1]; for brief summaries see Uher, 2018a, 2021b, pp. 219–222, 2022b, pp. 4–5). It provides overarching philosophical, metatheoretical and methodological frameworks that coherently build upon each other and into which established concepts from various disciplines have been systematically integrated and complemented by novel ones. These involve in particular (1) concepts of psyche, behaviour, language and contexts (e.g., Uher, 2013, 2016a, 2016b); (2) concepts and methodologies for taxonomising and comparing individual differences in various kinds of phenomena within and across populations (e.g., Uher, 2015a, 2015b, 2015c, 2015e, 2018b; Uher, Addessi, & Visalberghi, 2013), as well as (3) concepts and theories of data generation, quantification and measurement across the sciences (e.g., Uher, 2019, 2020b, 2022a) and in quantitative psychology and psychometrics (e.g., Uher, 2018a, 2021c, 2021d, 2022b; Uher & Visalberghi, 2016; Uher, Werner, & Gosselt, 2013). The TPS-Paradigm's frameworks therefore provide strong conceptual foundations for scrutinising rating 'scales' and the common assumptions that they could enable psychological 'measurement' as will be shown now.

2 | RATING 'SCALES': EFFICIENT TOOLS ENABLING PSYCHOLOGICAL 'MEASUREMENT'?

Like other empirical scientists, psychologists aim to explore their study phenomena by generating data about them, analysing these data and drawing inferences from the results to their study phenomena. These aims are undisputed. But their implementation in psychology is complicated.

2.1 | Psychology's key challenges

Conceptualising, analysing and interpreting are key scientific activities—and abilities of human minds. By studying minds, psychologists explore the very means by which science is made. This has intricate implications that are often overlooked. Four challenges are key.

2.1.1 | Challenge (1). Psychologists must clearly distinguish the study phenomena from the means of their exploration—The psych*ical* (mental) from the psych*ological*

Science, unlike non-scientific knowledge generation, provides ways of thinking simultaneously about phenomena *and* the means of producing knowledge about them. This requires (1) *metatheory*—philosophical and theoretical assumptions about the study phenomena's nature and the questions we can ask about them—and (2) *methodology*[2]—philosophical and theoretical assumptions about the ways that are suited for answering these questions (*approaches*) and the therefore useable procedures, operations and techniques (*methods*; Althusser & Balibar, 1970; Toomela, 2011).

Scientific activities like analysing, categorising and conceptualising are abilities of the human mind and empirical research is, by definition, experience-based (from Greek *empeiria* for experience). Thus, when studying mind and experience, psychologists' study phenomena are of the same kind as their means of exploring these phenomena. This entails intricate challenges because it complicates the logical distinction between the phenomena under study (e.g., experiences, intellectual abilities, beliefs) and the means for exploring these phenomena (e.g., terms, data, scientific constructs). This requires that researchers critically reflect on and explicate their philosophical and theoretical (pre)assumptions—and use a clear terminology. Terming the *phenomena of the psyche*[3] *in themselves* as 'psych*ical'* (e.g., mental, experiential; caution: not to be mistaken for paranormal or spiritualist) and the means of their exploration as 'psych*ological'* (from Greek *logos* for body of knowledge), as in many non-English languages,[4] reflect this vital distinction (Figure 1a; Lewin, 1936; Uher, 2016a). For example, a psychological problem is a professional problem of the scientific discipline; a psychical problem is one of individuals' mental health. The frequent English-language use of 'psychological' for both cannot make this distinction. Analogously, we get viral (not virological) infections but we do virological research. Failure to make this important logical distinction of the psychical from the psychological—called *psychologists' cardinal error* (Uher, 2022b)—is widely reflected in common psychological jargon and practices.

Without clear conceptual and terminological distinctions, psychologists are prone to conflate their means of exploration with the phenomena being explored. This problem is most difficult to identify in research on psychical phenomena, but it occurs in research on all study phenomena (e.g., behavioural and social phenomena), especially

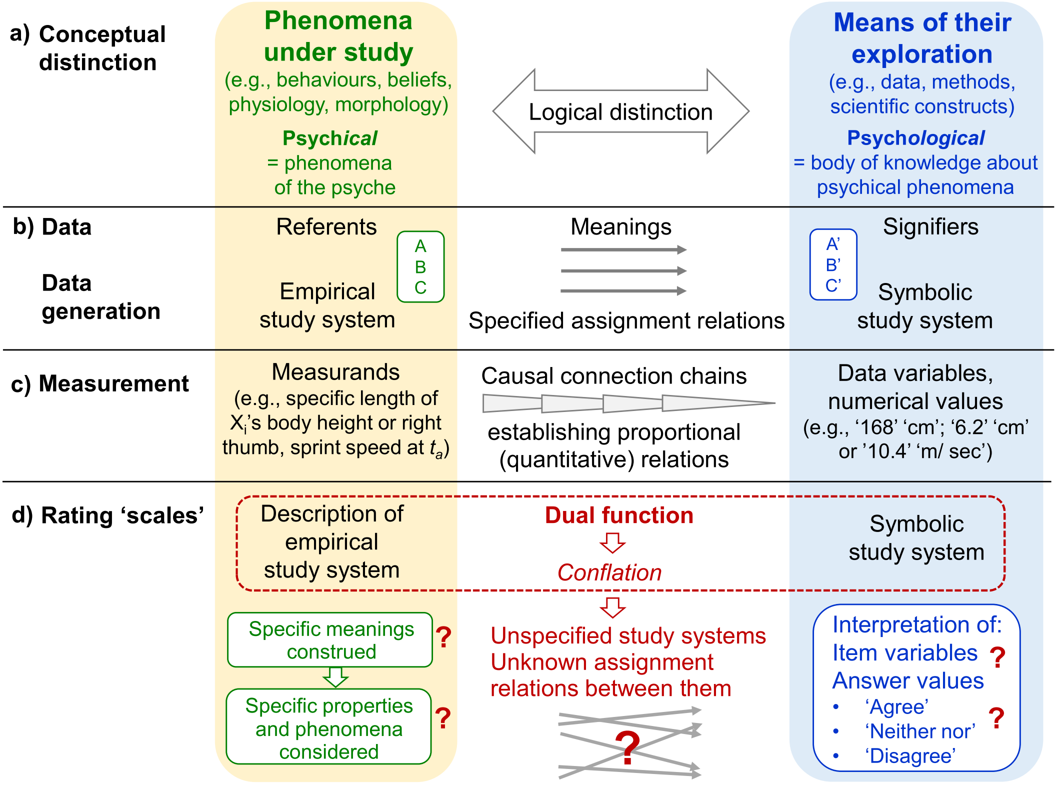
a) Conceptual distinction
Phenomena under study
(e.g., behaviours, beliefs, physiology, morphology)
Psych*ical*
= phenomena of the psyche
Logical distinction
Means of their exploration
(e.g., data, methods, scientific constructs)
Psycho*logical*
= body of knowledge about psychical phenomena
b) Data
Data generation
Referents
A
B
C
Empirical study system
Meanings
Specified assignment relations
Signifiers
A'
B'
C'
Symbolic study system
c) Measurement
Measurands
(e.g., specific length of X_i's body height or right thumb, sprint speed at t_a)
Causal connection chains
establishing proportional (quantitative) relations
Data variables, numerical values
(e.g., '168' 'cm'; '6.2' 'cm' or '10.4' 'm/ sec')
d) Rating 'scales'
Description of empirical study system
Dual function
Conflation
Symbolic study system
Specific meanings construed
?
Specific properties and phenomena considered
?
Unspecified study systems
Unknown assignment relations between them
?
Interpretation of:
Item variables ?
Answer values
• 'Agree'
• 'Neither nor'
• 'Disagree'
?

FIGURE 1 Distinction of the study phenomena from the means of their exploration: Relations to the basic principles of data generation and measurement and its failed implementation in rating 'scale' methods. (a) A key challenge for psychologists lies in the distinction of their study phenomena from their means of exploring these phenomena. This distinction is particularly challenging with regard to the phenomena of the psyche. This requires a precise terminology distinguishing the psychical (e.g., mental, experiential) from the psychological. The logical distinction between study phenomena and study means underlies the (b) conceptual components of semiotic systems like data as well as the key principles of data generation in general and of (c) measurement in particular. By contrast, the (d) dual function of rating 'scales' blurs this crucial distinction, leading to the frequent conflation of the study phenomena with the means of their exploration—psychologists' cardinal error—such as during data generation, analysis and interpretation.

in construct-based research. A construct is 'a *conceptual system* that *refers to* a set of entities—the construct *referents*—that are regarded as meaningfully related in some ways or for some purpose *although they actually never occur all at once* and that are therefore considered only on more abstract levels as a joint entity' (Uher, 2022b, p. 14). Thus, constructs do not exist as concrete entities in themselves; they are only thought of as entities—they are *conceptual entities* construed to efficiently refer to specific sets of referents. In everyday life and in science, we construe constructs (e.g., 'neuroticism'; 'environment', 'peace') about all kinds of phenomena (e.g., abiotic, biotic, psychical, social, cultural) and we tend to mistake the constructs for the phenomena to which they refer. This *construct–referent conflation* (Maraun & Gabriel, 2013) entails the erroneous practice of mistaking scientific constructs, thus the means of exploration, for the phenomena to which they refer, thus the actual phenomena under study (e.g., 'trait' constructs for the behavioural, emotional and cognitive phenomena to which they refer; Uher, 2013, 2022b). Similarly, the item *variables* that researchers use to encode and analyse information about the study phenomena—the variables' *referents* (e.g., individuals' age, behaviours or beliefs)—are often interpreted as if they constituted these study phenomena in themselves. This *variable–referent conflation* often occurs when the study phenomena (located in the individuals studied; e.g., age, behaviours or beliefs) and the item variables (located on spreadsheet and subjected to statistical analysis) are both labelled as 'variables' (Danziger & Dzinas, 1997; Uher, 2021a, 2021c, 2021d).

Failure to logically distinguish the study phenomena from the study means entails the conflation of disparate scientific activities, thereby making their distinction technically impossible and distorting scientific concepts and procedures (Uher, 2022b). This is also related to and complicated by further challenges.

2.1.2 | Challenge (2). Psychologists cannot be independent of their objects of research

Psychologists are studying many phenomena that are important in everyday life and in individuals of (primarily) their own kind. Psychologists start researching these only *after* having acquired, in their pre-academic lives, a complex pertinent everyday psychology (Uher, 2011, 2013). Everyday knowledge and language are pre-structuring researchers' minds (Danziger, 1997; Smedslund, 2016; Valsiner, 2012). Vague definitions and inconsistent use of key terms and concepts in scientific psychology (e.g., 'mind', 'behaviour'; Zagaria et al., 2020) may therefore derive from researchers' intuitive reliance on their everyday psychology, leading to widespread jingle–jangle fallacies (same term denotes different concepts, and vice versa; Uher, 2013, 2021b). In consequence, psychologists' own experiences—as humans, members of particular communities, and as individuals—may (unintentionally) influence their scholarly thinking (Danziger, 1997; Weber, 1949). This may entail anthropo-centric, ethno-centric and ego-centric (type-I and type-II) biases, such as when researchers misattribute properties of their own ingroup to outgroups or ignore outgroup properties uncommon in their ingroup (Uher, 2013, 2015b, 2015c, 2020a).

2.1.3 | Challenge (3). Psychology's study phenomena are heterogeneous and complex

Psychologists study complex and heterogenous phenomena occurring in all areas of human life (e.g., biotic, cognitive, social, developmental, cultural-societal). Their systematic integration, made necessary by these phenomena's joint occurrence in the single individual as psychology's basic unit of analysis, entails unparalleled challenges because these phenomena require different epistemologies, approaches and methods of exploration (Uher, 2021b). Moreover, human individuals are agents who subjectively interpret and reflect on their world. They act intentionally, memorise and learn. This limits possibilities for controlled and identically repeated cause–effect experiments (Bandura, 1986; Cabell & Valsiner, 2014; Fahrenberg, 2013; Rotter, 1954; Smedslund, 2002, 2004; Uher, 2021a).

Challenges also arise from the fact that psychologists explore many phenomena that are processes by nature (e.g., experience, behaviour). At any given moment, only parts of a process exist. Processual phenomena can thus be conceived only by generalising and abstracting from their occurrences over time—using abstract concepts, such as *constructs*.[5] *Conceptual abstraction* allows us to filter information of complex phenomena and reduce their complexity by emphasising some of their aspects and deemphasising others (Whitehead, 1929), depending on their ascribed (ir)relevance for a given meaning or purpose (e.g., social valence, prediction). As humans, we all intuitively *construe* abstract ideas (constructs) to describe and predict regularities in our world. We try to integrate our personal (idiosyncratic) and socially shared (folk) constructs and organise them at different levels of abstraction, thereby developing *construct hierarchies* (Kelly, 1955). That is, constructs can refer also to other constructs representing their contents on higher levels of abstraction (e.g., a construct 'nervousness' may refer to more specific constructs, such as 'insecurity', 'alertness' and 'timidity', each of these may refer to even more specific constructs etc., which eventually are related to specific observable events, such as behaviours). This entails nested conceptual structures (symbolised by words) in which more abstract constructs 'inherit' the meanings and referents from the various more specific constructs that they comprise, leading to broad fields of meaning (Uher, 2021d, 2022b). When psychologists study people's socially shared (folk) constructs of 'personality' (Tellegen, 1993), they develop more complex and abstract scientific constructs *about* these everyday constructs (Uher, 2013, 2015c, 2016a). That is, constructs are important means of exploration. But in some studies, constructs are the phenomena under study in themselves. This complicates the distinction between these disparate elements of research—but, in a given study, the same construct logically cannot be both. The Big Five 'personality' constructs, for example, are scientific constructs used to summarise people's everyday constructs and should therefore not be confused with these latter (Uher, 2022b).

In sum, constructs contain complex conceptual structures and broad implicit fields of meaning, which are symbolised by their linguistic labels. This is where language comes into play.

2.1.4 | Challenge (4). Language is essential for science but entails intricacies often overlooked

Language is basic to human life—and to science. What cannot be described cannot be researched (Wittgenstein, 1922). Language is so deeply engrained in everyday thinking that we easily overlook its inherently symbolic and composite nature. However, what we write or say typically bears no inherent relations to the objects referred (e.g., resemblance[6]). We can use written and spoken words (*signifier;* e.g., the phoneme [tri:] or the grapheme 'TREE') to refer to something (*referent;* e.g., a tree outside) only through the meanings (*signified;* e.g., the idea of a woody plant) that we attribute to both (Figure 1b). The representational function of signs thus arises from signifier–referent–meaning interrelations that are only conceptual and established by sociocultural conventions (Danziger, 1997; Deutscher, 2006; Ogden & Richards, 1923; Uher, 2015b, 2015d, 2016b, 2021a).

The semiotic function of human language allows us to turn—on mere conceptual levels—perceivable properties (e.g., white) into hypothetical objects (e.g., 'whiteness'; Peirce, 1958/1902, CP 4.227). Through this purely semiotic (sign-based) *reification* (*objectification*), we can make perceivable properties conceptually independent of their embodied experience. These reified properties can then become objects of consideration in themselves (e.g., 'colour') and can be linked to other perceptions, objects and meanings (e.g., 'whiteness' as socio-political category). This allows us to mentally handle abstract ideas and to abstract them further. Hence, languages have words with concrete referents as well as abstract words referring to ideas and concepts that are only distant from immediate perception (Vygotsky, 1962) and that we cannot easily trace anymore to their formerly concrete references and contexts (Deutscher, 2006). Words thus carry meanings that vary across time and contexts and that are drawn from their logical connections with other words in a language's semantic space (e.g., visualised in semantic networks) and in the given sentence(s) used (Arnulf et al., 2014; Neuman et al., 2012).

In consequence, psychologists must be wary of mistaking linguistic abstractions (e.g., 'traits') for concrete objects (fallacy of misplaced concreteness; Whitehead, 1929), must carefully distinguish their study phenomena (e.g., psychical processes, everyday constructs) from their means of exploration (e.g., terms, scientific constructs; Figure 1a), and consider that words' context-dependent fields of meaning make language-based methods inherently interpretative. What does this mean for psychological 'measurement'?

2.2 | Foundations of measurement across the sciences

Many psychical phenomena are accessible in others only through language (e.g., attitudes, intentions, feelings). The idea to generate numerical data about them using well-structured verbal 'scales' therefore seemed to open up promising ways to capitalise on the precision and accuracy of quantitative data and their mathematical and statistical analysis.

Common psychological interpretations of rating data reflect ideas, which—regardless of inevitable differences in discipline-specific theories and practices—also underly metrological[7] frameworks of physical measurement. They can be formulated as two criteria that characterise—on an abstract, general level as the most basic common denominators across sciences—a data generation process as *measurement* (Uher, 2020b, 2021c, 2021d, 2022a).

2.2.1 | Criterion (1) Justified attribution of the results to the measurands

Measurement involves structured processes that justify the attribution of the generated results to the *measurands*—the specific entities to be measured (Mari et al., 2017), such as individual A's specific length of its body height or right thumb or its specific speed in a 100m sprint at time point t_a. This *ontological claim* reflects that measurement is aimed at obtaining information about the measurands and nothing else.

2.2.2 | Criterion (2) Public interpretability of the results' quantitative meaning

Measurement processes establish a shared understanding of the results' quantitative meaning with regard to the measurands (Maul et al., 2019), such as how long exactly '6.2' 'cm' of length is and exactly how fast '37.58' 'km/h' is. This *semiotic claim* refers to the inherently symbolic function of data—they can serve their purposes only if their meaning is unambiguous and made transparent.

These criteria inform two distinct yet interrelated methodological principles for establishing data generation processes that enable measurement and for distinguishing them from other processes that do not (e.g., opinions, judgements).

2.2.3 | Principle (1) Data generation traceability: Establishing causal measurand–result connections

For justified attributions, the entire data generation process—from the measurands (the specific entities to be measured, e.g., in an individual) up to the results assigned to them—must be fully transparent and traceable. This requires operational structures, often implemented through measuring *instruments*, that *1a)* enable an empirical interaction with the measurand and *1b)* establish proportional (quantitative) relations between the measurand and the result assigned to it (Figure 1c). Given that many measurands are accessible only indirectly, establishing causal measurand–result connections often requires sequential empirical interactions between different properties, whereby the result of each interaction step depends on the result of the previous, such as in indirect measurement. For example, measuring an object's weight with a spring scale involves stepwise connections of its specific mass with >> gravity force >> length of spring deflection (each through physical laws) >> length of extension over measurement scale (through visual comparison) and the latter finally with >> the numerical values assigned as results (through semiotic encoding). Unbroken documented connection chains allow a result to be traced, in the inverse direction, back to the measurand, thus making the *entire data generation process* transparent and reproducible (Figure 2a; Uher, 2018a, 2020b).

2.2.4 | Principle (2) Numerical traceability: Establishing known quantity–result connections

But which quantity value should be assigned to a measurand and why? How do we know exactly how long '1' 'meter' is? We know this because, for physical properties, scientists *2a)* agreed on (initially often arbitrarily[8]) defined primary reference quantities (e.g., the prototype meter), which also define measurement units (e.g., the 'meter'), and they *2b)* codified all established reference quantities and their empirical interrelations (e.g., '1' 'meter' = '39.3701' 'inch'). Moreover, to ensure that numerical results have the same quantitative meaning across time and contexts (e.g., specific length of '1' 'meter'), scientists *2c)* empirically connected each primary quantity reference with all pertinent working references

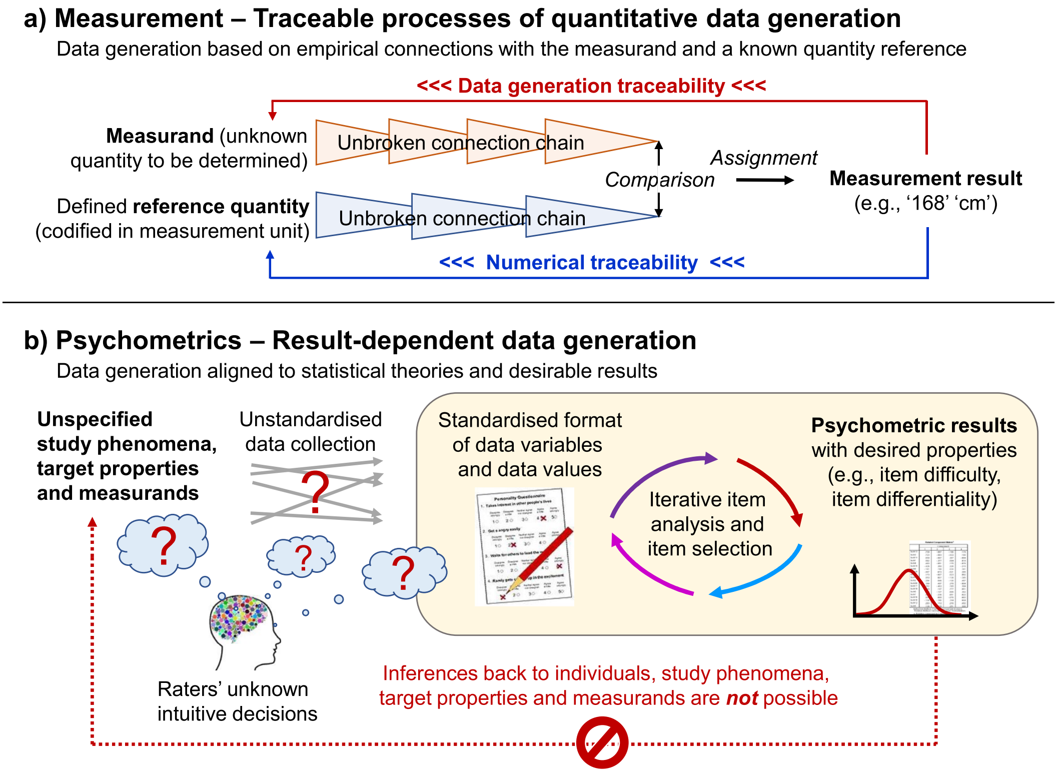
a) Measurement – Traceable processes of quantitative data generation
Data generation based on empirical connections with the measurand and a known quantity reference
<<< Data generation traceability <<<
Measurand (unknown quantity to be determined)
Unbroken connection chain
Defined reference quantity (codified in measurement unit)
Unbroken connection chain
Comparison
Assignment
Measurement result (e.g., '168' 'cm')
<<< Numerical traceability <<<
b) Psychometrics – Result-dependent data generation
Data generation aligned to statistical theories and desirable results
Unspecified study phenomena, target properties and measurands
Unstandardised data collection
?
?
?
?
Raters' unknown intuitive decisions
Standardised format of data variables and data values
Personality Questionnaire
Iterative item analysis and item selection
Psychometric results with desired properties (e.g., item difficulty, item differentiality)
Inferences back to individuals, study phenomena, target properties and measurands are not possible

FIGURE 2 Measurement versus psychometrics. (a) Measurement requires documented, unbroken chains of connections that establish proportional (quantitative) relations of the results with both the measurand's unknown quantity (e.g., person A's body height; principle of data generation traceability) and a known quantity reference (e.g., the international standard meter; principle of numerical traceability). Implemented in data generation processes, these basic methodological principles enable (1) the justified attribution of the results to the measurands (e.g., person A's body height) and (2) the public interpretability of the numerical result's quantitative meaning (e.g., how tall that is)—the two most basic criteria of measurement considered across the sciences. (b) In psychometrics, by contrast, item analysis and selection are used to develop instruments enabling the generation of data that differentiate well and consistently between cases (discrimination and reliability) and in ways considered important (validity). These between-individual analyses are necessary to first create quantitative meaning for rating scores that are, given the numerical recoding of verbal answer categories, devoid of information regarding the specific target properties, measurands and particular quantities to which they refer.

that are used for measurement execution (e.g., desk rulers) through world-wide networks of unbroken documented calibration[9] chains. These networks allow a result to be traced, in the inverse direction, back to a conventionally agreed and defined quantity reference, thus making the *result's quantitative meaning* transparent and publicly interpretable (Figure 2a). Notwithstanding inevitable and necessary differences, conventionally agreed quantitative meanings are also established in psychology, such as in time-based measurements of behavioural performances, physiological measurements and counts of test responses of defined correctness (e.g., in attention or achievement tests; Uher, 2021a).

These two methodological principles guide—on an abstract, general level—the necessary adaptations to the peculiarities of the disciplines' different study phenomena (which must consider further key elements of measurement, e.g., uncertainty and error; Giordani & Mari, 2014). Their implementation in discipline-specific theories and practices allows the generated results to be traced back to both (1) the measurands and (2) the quantity references used to determine the numerical values to be conventionally assigned to them, thereby making the entire measurement process *and* the results' quantitative meaning transparent and reproducible (Figure 2a; Uher, 2018a, 2020b, 2021d, 2022a).

In psychology, however, these principles are largely unknown. Although fundamental to measurement, most psychologists seem to be unfamiliar with the concept of the measurand. Indeed, psychologists rarely ever define the entities that they aim to measure in their studies as well as the quantities that they state to have 'measured' and indicated by the numerical values assigned during data generation. This highlights fundamental differences between *measurement* and psychologists' common practices for generating quantitative data with rating methods as illustrated now.

2.3 | What—Or who—Is the rating instrument interacting with the study phenomena?

Psychologists consider rating 'scales' as 'measuring' instruments that enable interactions with the study phenomena as well as standardised scoring. Yet it is the raters who must understand the verbal 'scales', use the meanings that these 'scales' have for them to identify specific relevant phenomena, and who must interact with these phenomena to generate data about them. That is, rating instruments inherently rely on human abilities, whereas technical instruments are designed to reduce human involvement in measurement processes (Uher, 2018a, 2019, 2020b).

The notion of rating 'scales' as 'measuring' instruments shifted psychologists' focus away from the persons interacting with both the verbal 'scales' and the study phenomena. Instead, psychologists' efforts for 'instrument development' are centred on the psychometric properties of the data that can be produced with rating 'scales', whereas the complex interactions executed by raters to generate these data in the first place remained largely unexplored. This also led psychologists to overlook serious methodological problems.

2.4 | The dual function of rating 'scales' masks key elements of measurement

Rating 'scales' serve two purposes. (1) They describe the phenomena and properties of interest (e.g., specific behaviours and their intensity; located in individuals). At the same time, (2) rating items and answer categories also serve as data variables and values (located on spreadsheet). That is, rating 'scales' function as both (1) description of the empirical study system and (2) symbolic study system used to explore that empirical system. This dual function may seem efficient, but it masks the crucial distinction between study phenomena and the means of exploration, thereby promoting their frequent conflation (psychologists' cardinal error; Figure 1d) and blurring disparate research activities (e.g., the definition of study phenomena with their empirical investigation; Uher, 2018a, 2021d, 2021c, 2022b).

This also obscures key elements of data generation, especially of measurement. Data[10] are sign systems that scientists use to encode information about their study phenomena. As signs, data can be stored, manipulated, decomposed and recomposed, thus analysed *in lieu of* the actual study phenomena and in ways not feasible for these phenomena in themselves. But inferences from the results back to the study phenomena can be made only if the data systems appropriately reflect relevant properties of these phenomena. This presupposes that the processes by which information from the empirical phenomena is encoded into the data—that is, *by which measurands are causally connected with the results*—are made transparent, reproducible, and thus traceable (Uher, 2020b).

Researchers must therefore specify (1) the system of the empirical phenomena studied (e.g., the specific behaviours studied), (2) the symbolic study system used to encode and analyse information about the empirical study system (e.g., the specific variables and values on spreadsheet), as well as (3) determinative assignment relations between these two study systems so that the same symbol always encodes the same information about the elements of the empirical study system (e.g., during observation). Put in semiotic terms, researchers must specify for their data systems (1) the referents, (2) the signifiers and (3) the meanings attributed to and thus linking the two former (Figure 1b; Uher, 2018a, 2021a).

This idea also underlies representational theory of measurement, developed in the social sciences (Krantz et al., 1971). It formalises, in representation theorems, axiomatic conditions for the mapping relations between the empirical and the symbolic relational system. Axiomatic conditions for the permissible transformations of the symbolic relational system without breaking its relations to the empirical relational system are formulated in uniqueness theorems (Vessonen, 2017). Psychologists are well-familiar with uniqueness theorems (e.g., for selecting statistical tests that are appropriate for given data types). But they often overlook (e.g., Borsboom & Mellenbergh, 2004) that explicit representation theorems are essential first steps[11] for implementing data generation traceability and numerical traceability in measurement processes (Uher, 2018a, 2020b, 2021c, 2021d). This methodological necessity is obscured by the dual function of rating 'scales', which entails that researchers specify neither the empirical nor the symbolic system nor, in consequence, the assignment relations between them (Figure 1d). This important task is left to raters.

2.5 | Intricate demands imposed on raters

Raters must, first, interpret the rating items and answer categories to identify relevant phenomena to be judged (e.g., specific behaviours) and the kind of grading enquired (e.g., frequencies, agreement). To rate how a person typically 'is' (e.g., in 'personality' ratings) or how intensely a feeling is perceived at a specific moment (e.g., in 'momentary' ratings), raters must also identify suitable references for comparison (e.g., other occasions). They must draw all this information from the rating instruments' colloquial wordings using their common-sense knowledge. How do raters do this?

Common-sense categories are often fuzzy and context-sensitive with flexible boundaries (Hammersley, 2013). Colloquially worded rating items (e.g., 'gets nervous easily') can therefore refer to broad ranges of phenomena, individuals and context, without specifying any particular ones. This contrasts with behavioural measurement, in which specific, physically described and situationally located behavioural acts (e.g., fidgeting, finger-tapping) are recorded in their occurrences over time (e.g., durations, frequencies). Ratings go far beyond this and typically enquire about an individual's intentions, abilities, feelings, etc., which can be inferred from behaviours but are not contained in behaviours themselves. Most behaviours are inherently ambiguous because they simultaneously possess various features and can therefore evoke different meanings (Shweder, 1977).

Meanings do not exist in themselves but always *for* someone. Therefore, meanings can be constructed from many possible interpretive viewpoints explaining behaviours, for example, by reference to intentions, goals, rules, situations or person characteristics ('traits'). Each of these interpretive perspectives follows logical principles (Kelly, 1955; Smedslund, 2002, 2004). But which particular perspective a person considers at any given moment is never logically determined by a behaviour itself (Shweder, 1977).

Fields of meaning that prove to be viable for 'reading' individuals' behaviour in everyday life become anchored in people's personal and socially shared everyday concepts (Kelly, 1955). Common-sense constructs that are viable for predicting and controlling individual behaviours—thus for differentiating individuals and establishing normativity—become encoded in person-descriptive everyday words[12] (Klages, 1926). More abstract words cover more diverse interpretive perspectives and have larger networks of logical connections with other words in a language's semantic space. Colloquially worded rating items therefore reflect whole networks of conventionally established meaning relations and interpretive possibilities and are thus inherently inferential (Arnulf et al., 2014; Block, 2010; Rosenbaum & Valsiner, 2011; Shweder & D'Andrade, 1980; Smedslund, 2002).

In consequence, raters must use their semantic knowledge and decide which particular item meanings to construct for a given rating. From this, raters must decide which behaviours and which of their possible interpretations to consider in order to infer specific phenomena to be judged (Figure 1d). Ratings often require raters to implicitly compare occurrences of the target phenomena between individuals and over time (e.g., 'personality' ratings). Past experiences, however, are memorised only in abstracted and conceptually integrated forms (Valsiner, 2012). Thus, raters must also use the beliefs and ideas that they have developed *about* the target phenomena *in general*. This explains why ratings can be made *on demand and even in absence* of the target phenomena (e.g., habitual—i.e., past—behaviours), thus *retrospectively* (Uher, 2016a). This is impossible for measurement because it requires an empirical interaction with the measurand (the entity to be measured), which therefore cannot be absent during data generation (Uher, 2019).

Hence, considerable demands are placed on raters. This may explain their frequent use of mental shortcuts, such as by relying on semantic similarity, common stereotypes or answer tendencies (Arnulf et al., 2014; Shweder, 1977; Uher, 2018a; Uher, Werner, & Gosselt, 2013; Wood et al., 2012), which entail countless well-described rater biases (Podsakoff et al., 2003; Tourangeau et al., 2000).

2.6 | Variations in item interpretation preclude traceability

Many 'personality' researchers regard the socially shared fields of meaning of broadly worded rating items (e.g., trait-adjectives like 'nervous') as useful for covering more diverse aspects of their constructs and for predicting broader ranges of behaviours though less specifically (known as fidelity–bandwidth trade-off in personality psychology; Borkenau & Müller, 1991). Figure 3, for example, depicts the disparate meanings that 112 raters construed in their interpretations (in open answer format) of the item 'gets nervous easily', operationalizing 'neuroticism' in a popular 'personality' inventory (BFI-10[13]; Uher & Dharyial, unpublished). Such broad fields of meaning reflect the viability of raters' *collective* knowledge for predicting behaviours for everyday purposes and are therefore useful for the sample-level analyses prevailing in psychology. However, raters' *collectively* considered field of meaning need not be congruent with that intended by researchers, even if all reliability and validity criteria are fulfilled (see, e.g., Arnulf et al., 2020; Uher & Visalberghi, 2016).

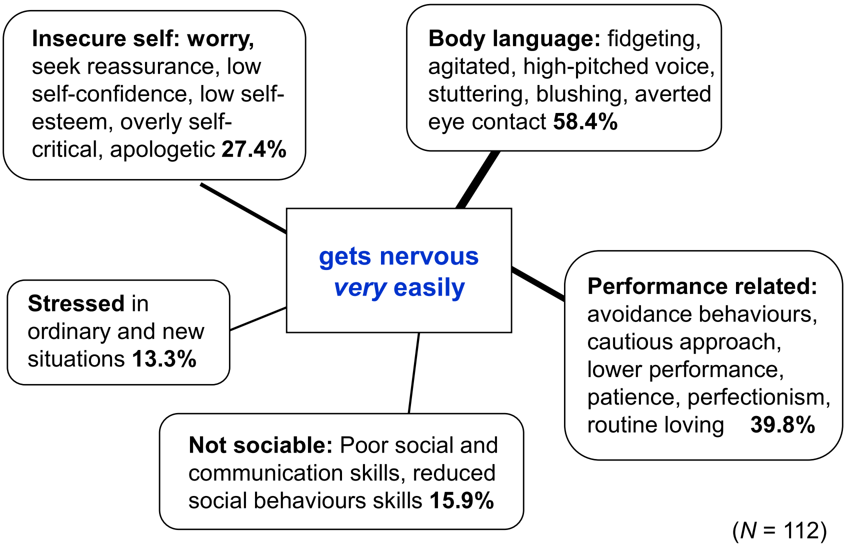
Insecure self: worry, seek reassurance, low self-confidence, low self-esteem, overly self-critical, apologetic 27.4%
Body language: fidgeting, agitated, high-pitched voice, stuttering, blushing, averted eye contact 58.4%
gets nervous very easily
Stressed in ordinary and new situations 13.3%
Performance related: avoidance behaviours, cautious approach, lower performance, patience, perfectionism, routine loving 39.8%
Not sociable: Poor social and communication skills, reduced social behaviours skills 15.9%
(N = 112)

FIGURE 3 Rating items: Broad semantic fields of meaning. Field of meaning of the item 'gets nervous easily', operationalising the 'personality' construct 'Neuroticism' in the BFI-10. The field is illustrated through the main themes that $N = 112$ raters mentioned in their open-ended item interpretations that they provided in terms of the behaviours that a fictitious target person scoring high on the item (indicated by 'very') would typically show. Percentages indicate the proportions of raters providing interpretations pertinent to a given theme (multiple nominations per person possible).

Yet for measurement, the interpretive flexibility of language entails serious problems. Not only can one and the same rating item always be interpreted differently—but also, for any given rating, each rater does not consider an item's *entire* field of *all* semantically possible meanings. Raters construe only *specific* meanings depending on the specific contexts they consider and raters differ in which ones they consider, both within and between individuals (Shweder, 1977; Smedslund, 2004; Uher, 2018a). For example, by considering the meaning of the item 'gets nervous easily' (Figure 3) as related to either body language, performance, self-concept or sociality; some raters may focus on fidgeting, others on perfectionistic behaviours and still others on reassurance seeking or communication skills. Different item interpretations entail that raters encode information about *different* empirical phenomena into the *same* symbolic element, making the ratings that are generated for the *same* item incomparable with one another (Figure 1d). These *many–to–one assignment relations* also preclude that the generated data can be traced back and thus be attributed to the measurands that raters may have considered—as this is required for measurement. The problems do not stop here; flexibly interpretable items also obscure the necessity to specify what is to be measured at all.

2.7 | Psychologists' neglect of the study properties and measurands

Psychologists' focus on colloquially worded items, enabling the efficient collection of overall judgements about a broad range of phenomena, led them to overlook that measurement requires specification of both the particular properties under study and the specific entities that are to be quantified—the measurands. So, what actually is it that psychologists aim to 'measure'?

The frequent yet erroneous interpretation of constructs as concrete real entities (reification), the equation of constructs with their referents (construct–referent conflation) and the common jargon of 'measuring constructs' led psychologists to overlook that constructs are abstract conceptual systems. Constructs are conceptual entities, which, consequently, cannot be measured in themselves. Inferential rating items like 'nervous' refer to sets of various concrete phenomena (see Figure 3) and thus describe abstract conceptual entities in themselves (Uher, 2018b). Common jargon (e.g., 'measuring behaviour') and the focus on verbal descriptions of the study phenomena in rating 'scales' also led psychologists to overlook that phenomena (and objects)—in themselves—cannot be measured either. From observational methods it is well-known that only specific behavioural acts shown by a specific individual at a particular time and place can be measured (e.g., individual A's sprint at time t_a), for example, regarding their temporal (e.g., duration) or spatial properties (e.g., length). Thus, researchers must specify which of the various *properties* that a phenomenon may feature they want to study (e.g., a specific gesture's spatial extension, execution speed or frequency of occurrence) and which of the various *entities of the given target property* that a phenomenon may feature—the *measurand*—they want to measure (e.g., the gesture's horizontal or vertical spatial extension shown by person P_a at time t_a; Allevard et al., 2005).

In rating 'scales', the property to be quantified is often indicated in the answer categories. This may apply to frequency 'scales' but what about the popular agreement (Likert) 'scales'? Can agreement reasonably be assumed to reflect quantities of phenomena as diverse as those subsumed as 'neuroticism', 'extraversion' and 'happiness'? Or does agreement not rather form part of the judgement process itself? Indeed, statistical findings are commonly *not* interpreted as reflecting raters' levels of agreement, as enquired during data generation, but instead as quantification results of the diverse phenomena *in themselves* that are described in the items (Uher, 2022a). This corresponds to re-interpreting a measurement result of, for example, length *at will* into one of mass, temperature or time—a practice that would never go unchallenged but that, in lack of specified study properties, goes undetected in psychology. Further problems derive from the kinds of 'scales' used in rating methods.

2.8 | Assigning graded judgements flexibly to a one–size–fits–all 'scale'

Raters must form and indicate their overall judgement using a bounded set of (mostly) verbal answer categories indicating staged degrees of the grading enquired (e.g., frequencies). These answer categories are commonly worded in abstract and general ways (e.g., 'seldom', 'sometimes', 'often') to enable applications to a broad range of phenomena and contexts. Raters must interpret these categories' meaning with regard to the specific item meaning and thus the specific phenomena, properties and measurands that they may consider in a rating. But how often is 'often' for a behaviour to occur given that occurrence rates generally vary between behaviours and across situations (e.g., talking vs. shouting; Uher, 2015a)? Regardless of the *different* phenomena that raters may consider for an item and that researchers enquire in *different* items, raters must always fit their judgements into the *same* set of answer categories provided in a rating 'scale'. That is, they must assign a broad range of quantitative information *flexibly* to a fixed, narrow range of values (e.g., five). This means, *raters must adapt their judgements to the 'scale' rather than to the phenomena to be judged*—and they can do so only by constructing *different* quantitative meanings for the *same* 'scale' category. This fundamentally contradicts the idea of measurement as enabling the accurate and reliable determination of quantities. Physical measurement scale units therefore have unchangeable quantitative meanings and pertinent values can be assigned to measurands without upper limits (Uher, 2022a).

To fit their judgements into the narrow bounded answer 'scale', raters sometimes—but not always—seem to intuitively weigh the study phenomena's observed occurrences against their presumed typical occurrence rates in given contexts (e.g., social groups), leading to well-known reference group effects (Heine et al., 2002; Uher, 2015a; Uher & Visalberghi, 2016; Uher, Werner, & Gosselt, 2013; Wood et al., 2012). Occurrences of individual behaviours are highly complex on all levels of consideration. Individuals differ in how they tend to behave in different situations (individual-specific situation–behaviour profiles) and in which behaviours of similar function (e.g., various aggressive acts) they tend to show in similar contexts (e.g., individual-specific response profiles). On the sample level, cross-situational and internal consistencies of variables encoding behaviours are therefore often only low to moderate (Asendorpf, 1988; Mischel et al., 2002; Uher, Addessi, & Visalberghi, 2013; Uher et al., 2008).

Our abilities to accurately track such complex occurrences are generally limited; therefore, raters may base their judgements more strongly on similarity in the behaviours' meanings (Shweder, 1977). But semantically guided judgements and the necessary flexible assignments can distort and even inverse quantitative relations (e.g., talking 'sometimes' may actually refer to more frequent occurrences than shouting 'often'), thereby introducing complex shifts in the quantitative meaning of the data produced (Uher, 2015a, 2022a; Uher, Werner, & Gosselt, 2013). Indeed, raters report very different reasons for choosing rating 'scale' boxes, which are often rather trivial and not even quantitative at all (Uher, 2018a), resulting in *many–to–one* and *one–to–many* relations with the 'scale' categories provided. Hence, raters interpret and use rating 'scales' not in standardised ways, as often assumed, but flexibly—as semantically and logically required. This entails, however, that the same result does not reflect the same information (Uher, 2018a, 2022a). Measurement, by contrast, requires *determinative one–to–one* assignments that encode the *same* quantities (e.g., same lengths) always in the *same* symbols so that results always represent the *same* quantitative information regarding the property studied. Thus, the symbolic study system must be mapped onto the empirical study system such that the created numerical structures appropriately represent the empirical structures observed (Figure 1b–1d; Ellis, 1966; Tal, 2020; Uher, 2018a). These fundamental differences in data generation with rating 'scales' versus measurement scales are also linked to different meanings of the term 'scale'.

2.9 | What actually is a 'scale'?

Psychological 'scales' are commonly referred to Stevens' (1946) four categories of variables (e.g., nominal, ordinal or interval) indicating that their numerical values represent information of different complexity (e.g., categorical or sequence information without or with equal intervals). These conceptual properties form part of the symbolic (data) system and determine the permissible transformations *during data analysis* that maintain its mapping relations with the empirical study system (thus determining the statistical tests applicable). This presupposes, however, that—*during data generation*—appropriate mapping relations have first been set up through traceable empirical connections that establish proportional (quantitative) relations between the measurand, a known quantity and the result (Figure 2a).

To achieve this, measurement scales have four different methodological functions. They serve as (1) *instruments* enabling empirical interactions with the measurand (e.g., weighing scale), and specify the (2) *structural data format* (e.g., numerical value plus measurement unit), the (3) *conceptual data format* ascribed to these structures (e.g., ratio scale) and the (4) *conventionally agreed reference quantity* used (e.g., kilogram scale). These four different functions are necessary at different stages of the measurement process and are therefore not interchangeable (Uher, 2022a). With rating 'scales', however, psychologists only implement a particular data format (e.g., variables with five possible values) and ascribe to these data particular conceptual properties (e.g., ordinality). But this neither enables the necessary empirical interaction with the measurand (data generation traceability) nor does it specify a reference quantity or at least the target property measured. These specifications, however, are necessary to determine the specific numerical value to be assigned as well as its quantitative meaning (numerical traceability).

2.10 | Numerical data need not have quantitative meaning

In rating methods, numerical data are created by recoding verbal answer categories into numerals[14] (e.g., '1', '2', '3', '4', '5'; response coding). Numerals are commonly interpreted as numbers; but they are not the same. Numerals (e.g., '5', 'V', '௫') are signifiers that are often used to mean numbers but that can also have other meanings (e.g., alphabetic letters, categories). Signifiers are invented and therefore vary arbitrarily as illustrated in Table 1, which shows the same data set in Arabic, Roman and Tamil numerals. Persons who are regularly using Arabic numerals to indicate numbers may readily ascribe to these numerals quantitative properties but may hesitate to do so when the same numerical data are depicted in the less familiar Roman and Tamil numerals. But these as well can have quantitative meanings—as much as all three types of numerals can be assigned just categorical, thus non-quantitative meanings.

By taking numerals for numbers (*numeral-number conflation*), psychologists ascribe to their numerically recoded rating data quantitative meanings (Uher, 2021c, 2021d). *Quantities* are divisible properties of the same kind, thus of the same *quality* (Hartmann, 1964), which feature particular testable relations (e.g., $2 < 3 < 4$) as specified in the axioms of quantity (e.g., equality, ordering, additivity; e.g., Barrett, 2003). Considering here just ordinal values for agreement 'scales', one could certainly say that 'strongly agree' ('5') indicates more agreement than 'agree' ('4'). But could 'agree' ('4')

TABLE 1 Data spreadsheet with Arabic, Roman and Tamil numerals.

Numerals	Arabic				Roman				Tamil			
Persons Variables	P_a	P_b	P_c	P_d	P_a	P_b	P_c	P_d	P_a	P_b	P_c	P_d
V_a	5	3	1	4	V	III	I	IV	௫	௩	௧	௪
V_b	4	4	2	3	IV	IV	II	III	௪	௪	௨	௩
V_c	3	2	3	5	III	II	III	V	௩	௨	௩	௫
V_d	5	3	2	1	V	III	II	I	௫	௩	௨	௧

Note: Various signs that can all have quantitative meanings but that are indicated with different signifiers—here Arabic, Roman and Tamil numerals. Habitual use can mislead to readily interpret particular numerals (e.g., Arabic numerals) only as numbers although they can also have just categorical meanings. Less familiar types of numerals (e.g., Roman and Tamil numerals) make this more directly apparent.

reflect more agreement than 'neither agree nor disagree' ('3'), often chosen to indicate 'inapplicable' (Uher, 2018a)? Does 'agree' ('4') really reflect more agreement than 'disagree' ('2')—or are agreeing and disagreeing with something not fundamentally different ideas? Semantically, two different qualities can be easily merged into one conceptual dimension (e.g., semantic differentials; Snider & Osgood, 1969). But what divisible properties could we identify in abstract concepts (e.g., 'nervousness') that refer to *different* phenomena each featuring qualitatively *different* properties (see Figure 3) as well as *different* quantities (e.g., intensities, frequencies, durations)? Rating scores are also often aggregated across items. But could answering *1x 'agree'* ('2') *and 1x 'disagree'* ('4'), thus, having a split opinion or inversed item interpretation, indicate (roughly) the *same* agreement (averaging '3') like answering *2x 'neither agree nor disagree'* ('3'), thus having 'no opinion'? The logico-semantic meanings of verbal answer categories—even if just ordinally conceived—are obviously discordant with the quantitative meanings that are commonly ascribed to their numerically recoded values (Uher, 2022a).

2.11 | Psychometrics: Adjusting data to statistical theories rather than to the study phenomena and properties

When scoring responses, psychologists actually do not assign numerical values *in relation to* a scale as is done in measurement. In measurement, the unit indicates a defined reference quantity (e.g., '1' 'cm' of length) that is used to determine the measurand's still unknown quantity (e.g., '3 cm'), which must therefore be of the same property as the reference quantity used (e.g., of length). Instead, in rating 'scales' psychologists *replace the verbal answer 'units' in themselves with numerals*, thereby creating 'scores' that are devoid of information regarding both the specific property studied (e.g., '3' *of what*—agreement, intensity, frequency or duration?) and the specific quantity of that property that these numerals are meant to indicate (e.g., *how much* of that is '3'?).

For such scores, quantitative meaning can be created only through between-case comparisons. Sample-level statistics are therefore necessary to first create quantitative meaning for numerically scored ratings (Uher, 2022b). But this approach fails if all cases score the same—a first indication of fundamental problems. Indeed, in lack of traceable connections between measurand, result and a known quantity, it means comparing scores with unknown quantity information in order to create quantitative meaning for them—a truly Münchhausenian[15] effort. In measurement, by contrast, the measurand's unknown quantity (e.g., person A's body height) is compared with that of a known reference quantity (e.g., the standard meter unit), which establishes the numerical result's quantitative meaning (e.g., how tall that is) and ensures its public interpretability (numerical traceability; Figure 2a; Uher, 2021d, 2021a, 2022a, 2022b).

To enable between-case comparisons, psychometricians develop instruments, such as rating 'scales', that allow the generation of data that differentiate well and consistently between cases (see e.g., Bandalos, 2018). This *result-dependent data generation*, however, aligns data generation and results to statistical criteria and theories rather than to the actual phenomena and properties studied (Figure 2b; Uher, 2021d). This psychometric practice has various serious implications that are still hardly considered. Only a few can be outlined here.

Inter-rater and internal reliabilities, for example, concern relations of between-case score differences between raters or between different variables on the sample level.[16] But these relations provide evidence of consistency neither in these scores' assignment to the measurands (data generation traceability) nor in the scores' proportional relations to known quantities—thus, in their quantitative meaning (numerical traceability) as required for measurement. This precludes these scores' quantitative interpretation and their justified attribution to the measurands as implied by the common jargon about 'rater accuracy' (Biesanz & Human, 2010; Kenny, 1991) or 'instrument validity' ('ability to measure what it purports to measure'; Kelley, 1927). In metrology, by contrast, *measurement accuracy* denotes how closely the determined value agrees with the measurand's true quantity value (e.g., a person's true core body temperature); *measurement precision* denotes closeness between values obtained in replicate measurements (e.g., temperature determined using the same and different thermometers; JCGM200:2012, 2012).

Measurement accuracy and precision thus concern the relations of results that can be generated for the *same* measurand of the *same* property. Psychological reliability concepts—theoretically—also concern relations between observed and (hypothetical) true scores or relations between repeated assessments. But they are compromised by the

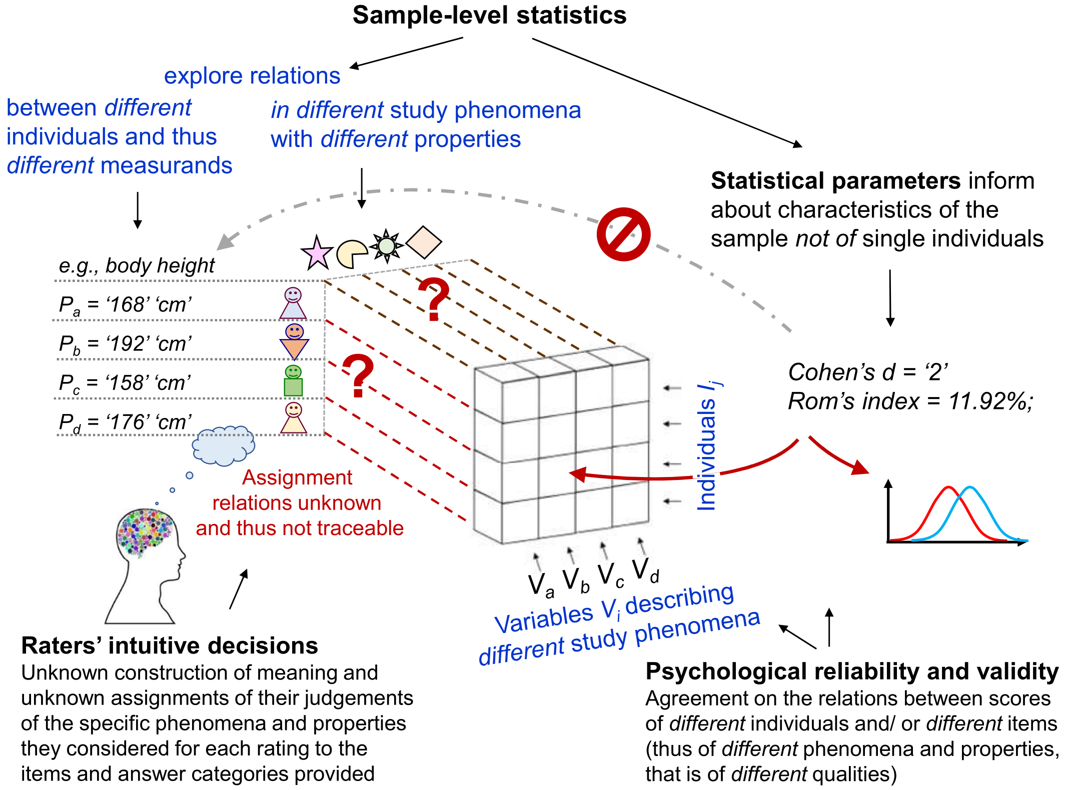
Sample-level statistics
explore relations
between different individuals and thus different measurands
in different study phenomena with different properties
Statistical parameters inform about characteristics of the sample not of single individuals
e.g., body height
P_a = '168' 'cm'
P_b = '192' 'cm'
P_c = '158' 'cm'
P_d = '176' 'cm'
?
?
Assignment relations unknown and thus not traceable
Individuals I_j
Cohen's d = '2'
Rom's index = 11.92%;
V_a V_b V_c V_d
Variables V_i describing different study phenomena
Raters' intuitive decisions
Unknown construction of meaning and unknown assignments of their judgements of the specific phenomena and properties they considered for each rating to the items and answer categories provided
Psychological reliability and validity
Agreement on the relations between scores of different individuals and/ or different items (thus of different phenomena and properties, that is of different qualities)

FIGURE 4 Sample-level statistics is not measurement. Psychological reliability and validity concepts and the therefore applied sample-level statistics concern relations between scores obtained for different individuals, thus different measurands, studied in often different phenomena with different properties (as intuitively considered by raters). Hence, they study relations between different qualities, whereas measurement is about capturing quantitative (divisible) properties of one specific quality. Moreover, sample-level results (e.g., effect sizes, correlations) are abstract parameters describing sample-level properties. They can reveal information about neither each given measurand's quantity (e.g., single individual's body height) nor about the quantitative meaning that the numerical result may have for the property studied (e.g., how tall is '168 cm'?). This precludes (1) the justified attribution of the results to the measurands (e.g., single individuals' body height) and (2) the public interpretability of the results' quantitative meaning—the two most basic criteria of measurement.

lack of causal measurand–result relations and the necessity of sample-level statistics for generating the scores' quantitative meaning, which analyse relations between scores obtained for *different* individuals and thus necessarily *different* measurands. Internal reliability and validity of rating 'scales', additionally, concern relations of scores obtained for *different* items and *different* constructs describing *different* phenomena with *different* properties (Figure 4; Uher, 2021c, 2021d). Indeed, psychological validity theories are about how particular study phenomena—such as those described as 'getting nervous easily' (Figure 2)—are related to other phenomena that are considered to be meaningfully related—such as those described as 'relaxed, handling stress well', used to study the internal validity of the BFI-10 'neuroticism' 'scale', or to later mental health problems that may be used to study its predictive validity. Hence, psychological validity concepts are about capturing relations between *different qualities*, whereas measurement is about capturing *quantitative* (divisible) properties of *one specific quality* (e.g., temperature). Consequently, summarising results obtained for different construct indicators[17] (even if these are results of measurement; e.g., durations of different behavioural acts) into construct indices (e.g., 'nervousness' index) constitutes not a step of measurement but one of (statistical) data modelling (Uher, 2020b, 2022b).

Psychologists' widespread use of sample-level statistics for studying individuals is based on the assumption that *inter*-individual and *intra*-individual variations are structurally identical (isomorph). Such isomorph structures are a property of stochastic processes and dynamic systems, called *ergodicity*, that fits all invariant phenomena (e.g., of non-living matter). But in phenomena undergoing change and development, such as those studied in psychology, ergodicity does not apply (Molenaar, 2008; Molenaar & Campbell, 2009; Valsiner, 2014). Thus, when assuming that *within*-individual structures could be explored by analysing *between*-individual structures, psychologists commit the *ergodic fallacy* (Speelman & McGann, 2020; van Geert, 2011). Prominent examples are the widespread equation of individual differences with 'personality' (Lamiell, 2013; Uher, 2018c) and the pervasive use of variable-oriented approaches (analysing score distributions across individuals from the viewpoint of single or multiple variables; Bergman & Trost, 2006; Stern, 1911) for studying individual functioning and development. But given ergodicity, sample-level findings can be generalised to individuals only if a) each individual obeys the same statistical model (homogeneity assumption) and if b) the statistical properties (e.g., factor loadings) are the same over time (stationarity assumption). These conditions, however, are rarely met in psychology (Molenaar, 2004; Molenaar & Campbell, 2009; Richters, 2021; Salvatore & Valsiner, 2010). Indeed, the idea that all individuals are the same (homogeneity assumption) fundamentally contradicts the concept of 'personality' (Uher, 2022b).

Numerical results obtained from sample-level analyses (e.g., correlations, effect sizes; Anvari & Lakens, 2021) may have well-established quantitative meanings in statistics. For example, effect sizes quantify the overlap between two groups' score distributions, such as men's and women's body height (e.g., Cohen's d = 0.3 corresponds to a Rom's non-overlap index of 11.92%; Rom & Hwang, 1996). But statistical scores are abstract parameters that describe distributions patterns in the sample and therefore inform neither about each given measurand's quantity (e.g., the single individuals' body height—data generation traceability) nor about the meaning of the particular quantity determined for a given measurand (e.g., *how tall* that is—numerical traceability; see Figure 4). This example of physical properties is obvious. But it highlights important points that are obscured in psychology by the lack of specified measurands and target properties and by the creation of numerical rating scores devoid of quantitative meaning regarding the property studied. Statistics neither is measurement (Fisher, 2009) nor is it therefore necessary. Measurement has been successful long before statistics was developed (Abran et al., 2012). Psychologists' reliance on sample-level statistics for the purposes of quantification is another key difference to measurement.

Some advocate for a 'soft' or 'wide' definition of measurement in psychology an social sciences (Finkelstein, 2003; Mari, 2013). Indeed, the level of measurement accuracy and precision, as necessary for sciences like physics, chemistry and medicine where errors can lead to the collapse of buildings, chemical explosions or drug overdoses, is not necessary for psychology. And yet, psychologists themselves often draw explicit analogies to physical measurement (e.g., in conjoint or Rasch measurement; Trendler, 2019, 2022; Uher, 2021c) and interpret even minor differences as meaningful. For example, in some countries, decisions on the death penalty for offenders partly rest on psychometrically determined IQ scores expressed to two-decimal place precision (Barrett, 2018)—although these are not results of accurate measurement but only pragmatic quantifications (Uher, 2021c) that require adjustment to be meaningful (Flynn, 2012; Young et al., 2007). It is only a matter of time before psychometric scores will be challenged in courts, like forensic psychologists' and psychiatrists' diagnostic practices before (Barrett, 2018; Faust, 2012).

Labelling different procedures uniformly as 'measurement' invites jingle fallacies (same term, different concepts) and misleads decision makers about the obtained scores' quantitative meaning and their justified attribution to the measurands (e.g., in individuals). Changing the definition of a key scientific activity cannot establish its comparability across sciences—it only undermines it (Uher, 2020b). Measurement is not just any activity for creating numerical data but involves structured documented processes that justify the high public trust placed in it (Abran et al., 2012; Porter, 1995). The principles of data generation traceability and numerical traceability provide guiding principles that specify—on the general methodological level—how measurement processes can be implemented in comparable ways in different sciences and be adapted to their study phenomena's peculiarities. These principles may also highlight inevitable limitations (Uher, 2020b, 2022a, 2022b).

3 | SO, WHAT IS WRONG WITH RATING 'SCALES'?

The alleged advantages of rating 'scales' for enabling efficient 'measurement' of psychical and behavioural phenomena using laypeople's judgements of brief everyday descriptions on one–size–fits–all 'scales' are also their greatest epistemological and methodological weaknesses. Ratings inherently rely on complex internal processes accessible only to raters. Unlike introspective methods, however, respondents are not given time to explore and explicate their experiences and interpretations of concrete events in the here and now. Instead, raters can reflect about the described phenomena only in abbreviated and decontextualised form and must intuitively make abstract, semantic and retrospective interpretations that remain unknown (Uher, 2016a; Valsiner, 2017).

Language is essential for psychological investigation, but its intricacies are often overlooked. The ease of rewording verbal materials to express any conceivable idea in various ways mislead many psychologists to assume that well-structured verbal descriptions could be used to generate numerical data that both meet desired statistical criteria and enable quantitative explorations of the phenomena of behaviour, psyche and society. With rating 'scales', psychologists run the risk of studying only linguistic propositions (Wittgenstein, 1922). These verbal 'scales' drew their attention to the conceptual-interpretive level and away from the actual study phenomena and their properties, which led them to overlook key elements of measurement (Uher, 2022b). Current proposals for improving psychological research practices involve, amongst others, more specific instructions to raters and administrators, items with narrower fields of meaning, and the provision of anchor-points for comparison (e.g., Anvari & Lakens, 2021). But these amendments cannot solve the problem that, in rating methods, a specific data point can be traced back neither to a concrete occurrence of a specific property in a concrete study phenomenon (data generation traceability) nor to a known quantity reference determining its quantitative meaning (numerical traceability), as is the case, by contrast, in behaviour observations. But both types of traceability are required to establish causal measurand–result connections in measurement and to ensure the quantitative results' public interpretability.

In pursuit of natural-science approaches, psychologists uncritically followed the promises of a seemingly quantitative method (see similarly, Debrouwere & Rosseel, 2022). The measurement jargon developed around rating 'scales' and psychometrics gave psychologists a false sense of advancement and of having established a solid scientific foundation. This misguided them to reduce efforts for method development and to shun critical reflection about the meaningfulness and interpretation of the numerical data produced and the quantitative analyses applied to them. Rating 'scales' do not enable implementation of unbroken connections that establish proportional relations of the results with both the measurands (data generation traceability) and known quantity references (numerical traceability) as required for measurement (Figure 2a). From rating scores, traceable connections cannot even be established to the verbal 'scales' that these scores numerically recode, not to mention the phenomena and properties that raters have considered in their ratings nor those that researchers have actually aimed to capture (Figure 2b). The necessity to create quantitative meaning for rating scores through between-individual comparisons mislead many psychologists to assume that quantitative results obtained from sample-level statistics could enable measurement (Figure 4). All this entails that attempts to infer possible quantitative structures in psychical phenomena from rating data are futile—regardless of whether or not one may find this quest meaningful.

The serious epistemological and methodological problems of rating 'scales' and the centrality of rating methods to many areas of psychological research highlight that just improving transparency in data analysis, as currently discussed, is insufficient for overcoming psychology's replication, confidence, validation and generalisability crises.

4 | FUTURE DIRECTIONS

To establish psychology as a science, psychologists must take intellectual responsibility for their discipline's philosophical and theoretical foundations (Uher, 2022b; Vautier et al., 2014). Scrutiny and controversial debate require critical self-reflection and making basic (implicit) assumptions explicit. This is a laborious task. It requires efforts for developing precise definitions, terminologies (Lilienfeld et al., 2015; Slaney & Garcia, 2015; Uher, 2021c, 2021d) and concepts (Bennett & Hacker, 2003; Uher, 2016b; Valsiner, 2021) that cannot build on everyday language. These may be more cumbersome and technical but are necessary to enable the crucial distinction between the study phenomena and the means of their exploration—thus, to avoid psychologists' cardinal error (Uher, 2022b). Acquiring some basic knowledge about semiotics and semantics can only be beneficial for mastering this challenging task (Valsiner, 2001; Vygotsky, 1962).

Ratings may be useful in applied fields, such as for opinion polling (e.g., about the frequencies of particular beliefs in a population). Psychometric approaches are useful for discriminating between responses in ways considered meaningful (e.g., social relevance). Such pragmatic, operationalist and instrumentalist approaches have their own justification and utility value and may be the most appropriate methods for studying many applied problems. But they preclude the scientific investigation of psychology's study phenomena, and thus its development as a science. These approaches must therefore be distinguished from the realist framework that many psychologists aim to pursue yet without implementing it also into their research practices (e.g., rating 'scales', psychometrics; Uher, 2018a, 2021c, 2021d). Psychologists need to expand their knowledge of methods and mathematics; and they can learn from the substantial advancements that other disciplines have made (Rudolph, 2013).

Meanwhile, rating methods inform a broad range of psychological research activities, including activities that should actually precede and not follow data generation. For example, many psychologists still consider only rating-based methodologies for developing 'personality' taxonomies (Condon et al., 2020) but ignore a broad range of alternative approaches that have meanwhile been developed to study, amongst others, variations in physiology and situated behaviours (Trofimova et al., 2018; Uher, 2008a, 2008b, 2013, 2015b, 2015c, 2015e). Such approaches are necessary to establish descriptive taxonomies of both the compositional structures and the process structures of specific kinds of phenomena as well as integrative and explanatory taxonomies involving various kinds of phenomena (Uher, 2018b).

The problems outlined in this article are just the tip of an iceberg of problems that—through the widespread and uncritical application of rating 'scales'—have become institutionalised in psychology (for details, see Uher, 2022b) and that therefore cannot be remedied with little quick fixes as many may hope. Elaborating future directions and possible solutions far exceeds the space of a journal article. Listing some here—such as ensuring that the study phenomena are actually present during data generation (e.g., as done in Ambulatory Assessment methods; Fahrenberg et al., 2007) because this is necessary to establish traceable relations to the results; studying the phenomena in context (e.g., Mehl, 2017) because this is necessary to explore their meanings and functions for individuals; developing theories about the study phenomena and properties in themselves rather than just relying on everyday beliefs about them because this is necessary to establish their quantitative and qualitative meanings (Uher, 2020b, 2022a); or replacing rating 'scales' with semantic computer algorithms to efficiently analyse open-ended verbal responses (Arnulf et al., 2021; Smedslund, 2021)—can only be incomplete. A first overview of future directions can be found in Uher (2022b), detailed elaborations will be published elsewhere (Uher, 2023).

This article, although perhaps unsettling for some, is meant to be a wake-up call, highlighting the necessity for urgent action and change of direction in ways that are still hardly considered in mainstream psychology. It would be desirable if this decade could see some of these changes happening.

ACKNOWLEDGEMENTS

This research was funded by a Marie Curie Fellowship of the European Commission's FP7 Programme awarded to the author (EC Grant Agreement number 629430). Many thanks to the editor and three reviewers for providing very detailed and constructive feedback.

CONFLICT OF INTEREST STATEMENT

The author declares to have no conflicts of interest.

ORCID

Jana Uher https://orcid.org/0000-0003-2450-4943

ENDNOTES

[1] http://researchonindividuals.org

[2] Methodology and method are not the same; methodology is the higher-order term. Their common conflation (e.g., by referring to methods as 'methodology', especially in English-language psychology) reflects many psychologists' reluctance to elaborate the philosophical and theoretical foundations of their research practices.

[3] The term psyche is conceived more broadly than mind, thus comprising non-mental phenomena as well.

[4] This distinction is made, for example, in French, Italian, Dutch, German and Russian.

[5] Constructs and concepts are both abstract ideas. Constructs tend to have more heterogeneous referents and therefore to be more abstract and complex. But attempts to clearly differentiate them are ultimately arbitrary.

[6] With very few exceptions (e.g., icons, onomatopoeia).

[7] Metrology, the science of measurement, foundational for the physical sciences and engineering.

[8] With modern technologies and using the knowledge gained from decreed measurement units, physicists replaced the originally arbitrary definitions of reference quantities with artefact-free definitions that are based on natural constants and are thus reproducible any time and any place (e.g., meter by speed of light; BIPM, 2006).

[9] These are called calibration chains because, along the connections in the chain, they specify uncertainties as a quantitative indication of measurement quality to assess a result's reliability and accuracy (JCGM100:2008, 2008).

[10] The term 'data' is also sometimes used to denote the study phenomena *in themselves*. This undifferentiated terminological usage promotes the conflation of the study phenomena with the study means (psychologists' cardinal error), which entails errors in conceptualisation and interpretation that often go unnoticed (Uher, 2022b).

[11] Representational theory of measurement provides, however, no concepts and procedures for implementing such theorems and it factors out important elements of measurement (e.g., measurement error and accuracy); therefore, it is insufficient as a theory of measurement in itself (Mari et al., 2017).

[12] This lexical hypothesis, first formulated by Galton, forms the basis of methodological approaches in which the person-descriptive words in a language's lexicon are used to categorise those individual differences that are considered most important in a given sociolinguistic community (e.g., Allport & Odbert, 1936; John et al., 1988; Uher, 2015c).

[13] Big Five 10-item short version (Rammstedt & John, 2007).

[14] Presenting raters with numerical rather than verbal answer categories does not solve the problems discussed here but only shifts the execution of this numerical recoding (response coding) to implicit considerations by raters.

[15] Referring to the famous story of Baron Münchhausen who pulled himself and the horse on which he was sitting out of a swamp by his own hair.

[16] This applies to the common variable-oriented approaches (studying score distributions across individuals). With individual-oriented approaches (studying within-individual score distributions across variables), reliability can be explored on the individual level yet regarding differences between scores obtained for *different* items, and thus *different* study phenomena and properties, rather than regarding the scores' relation to a given measurand's quantity.

[17] Given the complexity of constructs, only a (manageable) subset of a construct's *referents* can be empirically studied and are therefore chosen to serve as construct *indicators* (e.g., rating items; Uher, 2022b).

REFERENCES

Abran, A., Desharnais, J.-M., & Cuadrado-Gallego, J. J. (2012). Measurement and quantification are not the same: ISO 15939 and ISO 9126. *Journal of Software: Evolution and Process*, *24*(5), 585–601. https://doi.org/10.1002/smr.496

Allevard, T., Benoit, E., & Foulloy, L. (2005). Dynamic gesture recognition using signal processing based on fuzzy nominal scales. *Measurement*, *38*(4), 303–312. https://doi.org/10.1016/j.measurement.2005.09.007

Allport, G. W., & Odbert, H. S. (1936). Trait names: A psycholexical study. *Psychological Monographs*, *47*(1), 1–171. https://doi.org/10.1037/h0093360

Althusser, L., & Balibar, E. (1970). *Reading capital*. New Left Books.

Anderson, C. A., Allen, J. J., Plante, C., Quigley-McBride, A., Lovett, A., & Rokkum, J. N. (2018). The MTurkification of social and personality psychology. *Personality and Social Psychology Bulletin*, *45*(6), 842–850. https://doi.org/10.1177/0146167218798821

Andrade, C. (2021). HARKing, cherry-picking, p-hacking, fishing expeditions, and data dredging and mining as questionable research practices. *Journal of Clinical Psychiatry*, *82*(1). https://doi.org/10.4088/JCP.20f13804

Anvari, F., & Lakens, D. (2021). Using anchor-based methods to determine the smallest effect size of interest. *Journal of Experimental Social Psychology*, *96*, 104159. https://doi.org/10.1016/j.jesp.2021.104159

Arnulf, J. K., Larsen, K. R., Martinsen, Ø. L., & Bong, C. H. (2014). Predicting survey responses: How and why semantics shape survey statistics on organizational behaviour. *PLoS One*, *9*(9), e106361. https://doi.org/10.1371/journal.pone.0106361

Arnulf, J. K., Larsen, K. R., Martinsen, Ø. L., & Nimon, K. F. (2021). Editorial: Semantic algorithms in the assessment of attitudes and personality. *Frontiers in Psychology*, *12*, 720559. https://doi.org/10.3389/fpsyg.2021.720559

Arnulf, J. K., Nimon, K. F., Larsen, K. R., Hovland, C. V., & Arnesen, M. (2020). The priest, the sex worker, and the CEO: Measuring motivation by job type. *Frontiers in Psychology*, *11*, 1321. https://doi.org/10.3389/fpsyg.2020.01321

Asendorpf, J. B. (1988). Individual response profiles in the behavioral assessment of personality. *European Journal of Personality*, *2*(2), 155–167. https://doi.org/10.1002/per.2410020209

Bandalos, D. L. (2018). *Measurement theory and applications for the social sciences*. Guilford Publications.

Bandura, A. (1986). Social foundations of thought and action: A social cognitive theory. In *Social foundations of thought and action: A social cognitive theory*. Prentice-Hall, Inc.

Barrett, P. (2003). Beyond psychometrics. *Journal of Managerial Psychology*, *18*(5), 421–439. https://doi.org/10.1108/02683940310484026

Barrett, P. (2018). The EFPA test-review model: When good intentions meet a methodological thought disorder. *Behavioral Sciences*, *8*(1), 5. https://doi.org/10.3390/bs8010005

Bennett, M. R., & Hacker, P. M. S. (2003). *Philosophical foundations of neuroscience*. Wiley-Blackwell.

Bergman, L. R., & Trost, K. (2006). The person-oriented versus the variable-oriented approach: Are they complementary, opposites, or exploring different worlds? *Merrill-Palmer Quarterly*, *52*(3), 601–632. https://doi.org/10.1353/mpq.2006.0023

Biesanz, J. C., & Human, L. J. (2010). The cost of forming more accurate impressions: Accuracy-motivated perceivers see the personality of others more distinctively but less normatively than perceivers qithout an explicit goal. *Psychological Science*, *21*(4), 589–594. https://doi.org/10.1177/0956797610364121

BIPM. (2006). BIPM: The international system of units (SI) (8th ed.). Retrieved from http://www.bipm.org/

Block, J. (2010). The Five-Factor framing of personality and beyond: Some ruminations. *Psychological Inquiry*, *21*(1), 2–25. https://doi.org/10.1080/10478401003596626

Borkenau, P., & Müller, B. (1991). Breadth, bandwidth, and fidelity of personality-descriptive categories. *European Journal of Personality*, *5*(4), 309–322. https://doi.org/10.1002/per.2410050404

Borsboom, D., & Mellenbergh, G. J. (2004). Why psychometrics is not pathological. *Theory and Psychology*, *14*(1), 105–120. https://doi.org/10.1177/0959354304040200

Buntins, M., Buntins, K., & Eggert, F. (2016). Psychological tests from a (fuzzy-)logical point of view. *Quality and Quantity*, *50*(6), 2395–2416. https://doi.org/10.1007/s11135-015-0268-z

Cabell, K. R., & Valsiner, J. (2014). *The catalyzing mind. Beyond models of causality*. Springer. https://doi.org/10.1007/978-1-4614-8821-7

Condon, D., Beck, E., & Jackson, J. (2021). Age differences in personality structure. *Innovation in Aging*. *5*(Supplement_1), 564. https://doi.org/10.1093/geroni/igab046.2153

Condon, D., Wood, D., Mõttus, R., Booth, T., Costantini, G., Greiff, S., Johnson W., Lukaszewski A., Murray A., Revelle W., Wright A. G. C., Ziegler M., Zimmermann, J. (2020). Bottom-up construction of a personality taxonomy. *European Journal of Psychological Assessment*, *36*(6), 923–934. https://doi.org/10.1027/1015-5759/a000626

Danziger, K. (1985). The methodological imperative in psychology. *Philosophy of the Social Sciences*, *15*(1), 1–13. https://doi.org/10.1177/004839318501500101

Danziger, K. (1997). *Naming the mind: How psychology found its language*. Sage.

Danziger, K., & Dzinas, K. (1997). How psychology got its variables. *Canadian Psychology / Psychologie Canadienne*, *38*(1), 43–48. https://doi.org/10.1037/0708-5591.38.1.43

Debrouwere, S., & Rosseel, Y. (2022). The conceptual, cunning, and conclusive experiment in psychology. *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, *17*(3), 852–862. https://doi.org/10.1177/17456916211026947

Deutscher, G. (2006). *The unfolding of language: The evolution of mankind's greatest invention*. Arrow.

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00621

Ellis, B. (1966). *Basic concepts of measurement*. Cambridge University Press.

Fahrenberg, J. (2013). *Zur Kategorienlehre der Psychologie: Komplementaritätsprinzip; Perspektiven und Perspektiven-Wechsel On the category theory of psychology: Principle of complementarity, perspectives and changes of perspectives*. Pabst Science Publishers.

Fahrenberg, J., Myrtek, M., Pawlik, K., & Perrez, M. (2007). Ambulatory assessment - monitoring behavior in daily life settings. *European Journal of Psychological Assessment*, *23*(4), 206–213. https://doi.org/10.1027/1015-5759.23.4.206

Faust, D. (2012). *Ziskin's coping with psychiatric and psychological testimony*. Oxford University Press. https://doi.org/10.1093/med:psych/9780195174113.001.0001

Finkelstein, L. (2003). Widely, strongly and weakly defined measurement. *Measurement*, *34*(1), 39–48. https://doi.org/10.1016/s0263-2241(03)00018-6

Fisher, W. P. (2009). Invariance and traceability for measures of human, social, and natural capital: Theory and application. *Measurement*, *42*(9), 1278–1287. https://doi.org/10.1016/J.MEASUREMENT.2009.03.014

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–465. https://doi.org/10.1177/2515245920952393

Flynn, J. R. (2012). *Are we getting smarter? Rising IQ in the twenty-first century*. Cambridge University Press.

Giordani, A., & Mari, L. (2014). Modeling measurement: Error and uncertainty. In M. Boumans, G. Hon, & A. Peterson (Eds.), *Error and uncertainty in scientific practice* (pp. 79–96). Pickering & Chatto.

Haig, B. D., & Borsboom, D. (2008). On the conceptual foundations of psychological measurement. *Measurement: Interdisciplinary Research and Perspectives*, *6*(1–2), 1–6. https://doi.org/10.1080/15366360802035471

Hammersley, M. (2013). *The myth of research-based policy and practice*. Sage Publications Ltd. https://doi.org/10.4135/9781473957626

Hanel, P. H. P., & Vione, K. C. (2016). Do student samples provide an accurate estimate of the General Public? *PLoS One*, *11*(12), e0168354. https://doi.org/10.1371/journal.pone.0168354

Hardwicke, T. E., Thibault, R. T., Kosie, J. E., Wallach, J. D., Kidwell, M. C., & Ioannidis, J. P. A. (2022). Estimating the prevalence of transparency and reproducibility-related research practices in psychology (2014–2017). *Perspectives on Psychological Science*, *17*(1), 239–251. https://doi.org/10.1177/1745691620979806

Hartmann, N. (1964). *Der Aufbau der realen Welt. Grundriss der allgemeinen Kategorienlehre [The structure of the real world. Outline of the general theory of categories] 1940* (3rd ed.). Walter de Gruyter.

Heine, S. J., Lehman, D. R., Peng, K., & Greenholtz, J. (2002). What's wrong with cross-cultural comparisons of subjective Likert scales?: The reference-group effect. *Journal of Personality and Social Psychology*, *82*(6), 903–918. https://doi.org/10.1037/0022-3514.82.6.903

JCGM100:2008. (2008). Evaluation of measurement data – guide to the expression of uncertainty in measurement (GUM). Joint Committee for Guides in Metrology (originally published in 1993) Retrieved from http://www.bipm.org/en/publications/guides/gum.html

JCGM200:2012. (2012). *International vocabulary of metrology – basic and general concepts and associated terms* (3rd ed.). VIM. Retrieved from https://www.bipm.org/utils/common/documents/jcgm/JCGM_200_2012.pdf

John, O. P., Angleitner, A., & Ostendorf, F. (1988). The lexical approach to personality: A historical review of trait taxonomic research. *European Journal of Personality*, *2*(3), 171–203. https://doi.org/10.1002/per.2410020302

Kelley, T. L. (1927). *Interpretation of educational measurements*. World.

Kelly, G. (1955). *The psychology of personal constructs* (Vols. 1 & 2). Routledge.

Kenny, D. A. (1991). A general model of consensus and accuracy in interpersonal perception. *Psychological Review*, *98*(2), 155–163. https://doi.org/10.1037/0033-295x.98.2.155

Klages, L., & Johnston, W. H. (1926). Grundlagen der Charakterkunde [The science of character; Trans. 1932). Retrieved from http://dispater.atspace.com/

Krantz, D., Luce, R. D., Tversky, A., & Suppes, P. (1971). *Foundations of measurement Volume I: Additive and polynomial representations*. Academic Press.

Laajaj, R., Macours, K., Pinzon Hernandez, D. A., Arias, O., Gosling, S. D., Potter, J., Rubio-Codina, M., & Vakis, R. (2019). Challenges to capture the big five personality traits in non-WEIRD populations. *Science Advances*, *5*(7). eaaw5226. https://doi.org/10.1126/sciadv.aaw5226

Lamiell, J. (2013). Statisticism in personality psychologists' use of trait constructs: What is it? How was it contracted? Is there a cure? *New Ideas in Psychology*, *31*(1), 65–71. https://doi.org/10.1016/j.newideapsych.2011.02.009

Lamiell, J. (2019). *Psychology's misuse of statistics and persistent dismissal of its critics*. Palgrave Macmillan Cham. https://doi.org/10.1007/978-3-030-12131-0

Lewin, K. (1936). *Principles of topological psychology*. McGraw-Hill.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *22*(140), 1–55.

Lilienfeld, S. O., Sauvigné, K. C., Lynn, S. J., Cautin, R. L., Latzman, R. D., & Waldman, I. D. (2015). Fifty psychological and psychiatric terms to avoid: A list of inaccurate, misleading, misused, ambiguous, and logically confused words and phrases. *Frontiers in Psychology*, *6*, 1100. https://doi.org/10.3389/fpsyg.2015.01100

Ludeke, S. G., & Larsen, E. G. (2017). Problems with the big five assessment in the world values survey. *Personality and Individual Differences*, *112*, 103–105. https://doi.org/10.1016/j.paid.2017.02.042

Maraun, M. D., & Gabriel, S. M. (2013). Illegitimate concept equating in the partial fusion of construct validation theory and latent variable modeling. *New Ideas in Psychology*, *31*(1), 32–42. https://doi.org/10.1016/J.NEWIDEAPSYCH.2011.02.006

Mari, L. (2013). A quest for the definition of measurement. *Measurement*, *46*(8), 2889–2895. https://doi.org/10.1016/J.MEASUREMENT.2013.04.039

Mari, L., Carbone, P., Giordani, A., & Petri, D. (2017). A structural interpretation of measurement and some related epistemological issues. *Studies in History and Philosophy of Science*, *65–66*, 46–56. https://doi.org/10.1016/j.shpsa.2017.08.001

Maul, A., Mari, L., & Wilson, M. (2019). Intersubjectivity of measurement across the sciences. *Measurement*, *131*, 764–770. https://doi.org/10.1016/J.MEASUREMENT.2018.08.068

Mehl, M. R. (2017). The electronically activated recorder (EAR): A method for the naturalistic observation of daily social behavior. *Current Directions in Psychological Science*, *26*(2), 184–190. https://doi.org/10.1177/0963721416680611

Michell, J. (1999). Measurement in psychology. A critical history of a methodological concept. https://doi.org/10.1017/CBO9780511490040

Mischel, W., Shoda, Y., & Mendoza-Denton, R. (2002). Situation-behavior profiles as a locus of consistency in personality. *Current Directions in Psychological Science*, *11*(2), 50–54. https://doi.org/10.1111/1467-8721.00166

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement: Interdisciplinary Research & Perspective*, *2*(4), 201–218. https://doi.org/10.1207/s15366359mea0204_1

Molenaar, P. C. M. (2008). On the implications of the classical ergodic theorems: Analysis of developmental processes has to focus on intra-individual variation. *Developmental Psychobiology*, *50*(1), 60–69. https://doi.org/10.1002/dev.20262

Molenaar, P. C. M., & Campbell, C. G. (2009). The new person-specific paradigm in psychology. *Current Directions in Psychological Science*, *18*(2), 112–117. https://doi.org/10.1111/j.1467-8721.2009.01619.x

Neuman, Y., Turney, P., & Cohen, Y. (2012). How language enables abstraction: A study in computational cultural psychology. *Integrative Psychological and Behavioral Science*, *46*(2), 129–145. https://doi.org/10.1007/s12124-011-9165-8

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M, & Yarkoni, T. (2015). Promoting an open research culture: Author guidelines for journals could help to promote transparency, openness, and reproducibility. *Science (New York, N.Y.)*, *348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374

Ogden, C. K., & Richards, I. A. (1923). *The meaning of meaning: A study of the influence of language upon thought and of the science of symbolism*. Harcourt, Brace & World.

Peirce, C. S. (1958). *Collected papers of Charles Sanders Peirce, Vols. 1-6*. In C. Hartshorne, P. Weiss, & A. W. Burks (Eds.) (Vols. 7–8). Harvard University Press.

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*(5), 879–903. https://doi.org/10.1037/0021-9010.88.5.879

Porter, T. M. (1995). *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton University Press.

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the big five inventory in English and German. *Journal of Research in Personality*, *41*(1), 203–212. https://doi.org/10.1016/j.jrp.2006.02.001

Richters, J. E. (2021). Incredible utility: The lost causes and causal debris of psychological science. *Basic and Applied Social Psychology*, *43*(6), 366–405. https://doi.org/10.1080/01973533.2021.1979003

Rom, D. M., & Hwang, E. (1996). Testing for individual and population equivalence based on the proportion of similar responses. *Statistics in Medicine*, *15*(14), 1489–1505. https://doi.org/10.1002/(sici)1097-0258(19960730)15:14<1489::aid-sim293>3.0.co;2-s

Rosenbaum, P. J., & Valsiner, J. (2011). The un-making of a method: From rating scales to the study of psychological processes. *Theory & Psychology*, *21*(1), 47–65. https://doi.org/10.1177/0959354309352913

Rotter, J. B. (1954). Social learning and clinical psychology. In *Social learning and clinical psychology*. Prentice-Hall, Inc. https://doi.org/10.1037/10788-000

Rudolph, L. (2013). *Qualitative mathematics for the social sciences. Mathematical models for research on cultural dynamics*. In Routledge.

Salvatore, S., & Valsiner, J. (2010). Between the general and the unique: Overcoming the nomothetic versus idiographic opposition. *Theory & Psychology*, *20*(6), 817–833. https://doi.org/10.1177/0959354310381156

Shweder, R. A., Casagrande, J. B., Fiske, D. W., Greenstone, J. D., Heelas, P., & Lancy, D. F. (1977). Likeness and likelihood in everyday thought: Magical thinking in judgments about personality. *Current Anthropology*, *18*(4), 637–658. https://doi.org/10.1086/201974

Shweder, R. A., & D'Andrade, R. G. (1980). The systematic distortion hypothesis. In R. A. Shweder (Ed.), *Fallible judgment in behavioral research: New directions for methodology of social and behavioral science* (Vol. 4, pp. 37–58). Jossey-Bass.

Slaney, K. L., & Garcia, D. A. (2015). Constructing psychological objects: The rhetoric of constructs. *Journal of Theoretical & Philosophical Psychology*, *35*(4), 244–259. https://doi.org/10.1037/teo0000025

Smedslund, J. (2002). From hypothesis-testing psychology to procedure-testing psychologic. *Review of General Psychology*, *6*(1), 51–72. https://doi.org/10.1037/1089-2680.6.1.51

Smedslund, J. (2004). *Dialogues about a new psychology*. Taos Inst Publications.

Smedslund, J. (2016). Why psychology cannot be an empirical science. *Integrative Psychological and Behavioral Science*, *50*(2), 185–195. https://doi.org/10.1007/s12124-015-9339-x

Smedslund, J. (2021). From statistics to trust: Psychology in transition. *New Ideas in Psychology*, *61*, 100848. https://doi.org/10.1016/j.newideapsych.2020.100848

Snider, J. G., & Osgood, C. E. (1969). *Semantic differential technique: A sourcebook*. Aldine.

Speelman, C. P., & McGann, M. (2020). Statements about the pervasiveness of behavior require data about the pervasiveness of behavior. *Frontiers in Psychology*, *11*, 594675. https://doi.org/10.3389/fpsyg.2020.594675

Stern, W. (1911). *Die Differentielle Psychologie in ihren methodischen Grundlagen [Differential psychology in its methodological foundations]*. Barth.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 667–680. https://doi.org/10.1126/science.103.2684.677

Szollosi, A., Kellen, D., Navarro, D. J., Shiffrin, R., van Rooij, I., Van Zandt, T., & Donkin, C. (2020). Is preregistration worthwhile? *Trends in Cognitive Sciences*, *24*(2), 94–95. https://doi.org/10.1016/j.tics.2019.11.009

Tafreshi, D., Slaney, K. L., & Neufeld, S. D. (2016). Quantification in psychology: Critical analysis of an unreflective practice. *Journal of Theoretical & Philosophical Psychology*, *36*(4), 233–249. https://doi.org/10.1037/teo0000048

Tal, E. (2020). Measurement in science. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy (fall 2020)*. Retrieved from https://plato.stanford.edu/archives/fall2020/entries/measurement-science

Tellegen, A. (1993). Folk concepts and psychological concepts of personality and personality disorder. *Psychological Inquiry*, *4*(2), 122–130. https://doi.org/10.1207/s15327965pli0402_12

Teo, T. (2018). Outline of theoretical psychology. *Outline of theoretical psychology: Critical investigations. Palgrave Studies in the Theory and History of Psychology*. https://doi.org/10.1057/978-1-137-59651-2

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, *33*(4), 529–554. https://doi.org/10.1086/214483

Toomela, A. (2011). Travel into a fairy land: A critique of modern qualitative and mixed methods psychologies. *Integrative Psychological and Behavioral Science*, *45*(1), 21–47. https://doi.org/10.1007/s12124-010-9152-5

Toomela, A. (2018). *The psychology of scientific inquiry*. Springer International Publishing. https://doi.org/10.1007/978-3-030-31449-1

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge University Press. https://doi.org/10.1017/CBO9780511819322

Trendler, G. (2009). Measurement theory, psychology and the revolution that cannot happen. *Theory & Psychology*, *19*(5), 579–599. https://doi.org/10.1177/0959354309341926

Trendler, G. (2013). Measurement in psychology: A case ofignoramus et ignorabimus? A rejoinder. *Theory & Psychology*, *23*(5), 591–615. https://doi.org/10.1177/0959354313490451

Trendler, G. (2019). Conjoint measurement undone. *Theory and Psychology*, *29*(1), 100–128. https://doi.org/10.1177/0959354318788729

Trendler, G. (2022). The incoherence of Rasch measurement: A critical comparison between measurement in psychology and physics. *Personality and Individual Differences*, *189*, 111408. https://doi.org/10.1016/j.paid.2021.111408

Trofimova, I., Robbins, T. W., Sulis, W. H., & Uher, J. (2018). Taxonomies of psychological individual differences: Biological perspectives on millennia-long challenges. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1744), 20170152. https://doi.org/10.1098/rstb.2017.0152

Uher, J. (2008a). Comparative personality research: Methodological approaches. *European Journal of Personality*, *22*(5), 427–455. https://doi.org/10.1002/per.680

Uher, J. (2008b). Three methodological core issues of comparative personality research. *European Journal of Personality*, *22*(5), 475–496. https://doi.org/10.1002/per.688

Uher, J. (2011). Individual behavioral phenotypes: An integrative meta-theoretical framework. Why 'behavioral syndromes' are not analogs of 'personality. *Developmental Psychobiology*, *53*(6), 521–548. https://doi.org/10.1002/dev.20544

Uher, J. (2013). Personality psychology: Lexical approaches, assessment methods, and trait concepts reveal only half of the story-Why it is time for a paradigm shift. *Integrative Psychological and Behavioral Science*, *47*(1), 1–55. https://doi.org/10.1007/s12124-013-9230-6

Uher, J. (2015a). Comparing individuals within and across situations, groups and species: Metatheoretical and methodological foundations demonstrated in primate behaviour. In D. Emmans & A. Laihinen (Eds.), *Comparative neuropsychology and brain imaging*, *Series Neuropsychology: An interdisciplinary approach* (Vol. 2, pp. 223–284). Lit Verlag.

Uher, J. (2015b). Conceiving 'personality': Psychologist's challenges and basic fundamentals of the transdisciplinary philosophy-of-science paradigm for research on individuals. *Integrative Psychological and Behavioral Science*, *49*(3), 398–458. https://doi.org/10.1007/s12124-014-9283-1

Uher, J. (2015c). Developing 'personality' taxonomies: Metatheoretical and methodological rationales underlying selection approaches, methods of data generation and reduction principles. *Integrative Psychological and Behavioral Science*, *49*(4), 531–589. https://doi.org/10.1007/s12124-014-9280-4

Uher, J. (2015d). Agency enabled by the psyche: Explorations using the transdisciplinary philosophy-of-science paradigm for research on individuals. In C. W. Gruber, M. G. Clark, S. H. Klempe, & J. Valsiner (Eds.), *Constraints of agency: Explorations of theory in everyday life. Annals of Theoretical Psychology* (Vol. 12, pp. 177–228). Springer International Publishing. https://doi.org/10.1007/978-3-319-10130-9_13

Uher, J. (2015e). Interpreting 'personality' taxonomies: Why previous models cannot capture individual-specific experiencing, behaviour, functioning and development. Major taxonomic tasks still lay ahead. *Integrative Psychological and Behavioral Science*, *49*(4), 600–655. https://doi.org/10.1007/s12124-014-9281-3

Uher, J. (2016a). Exploring the workings of the Psyche: Metatheoretical and methodological foundations. In J. Valsiner, G. Marsico, N. Chaudhary, T. Sato, & V. Dazzani (Eds.), *Psychology as the science of human being: The Yokohama Manifesto* (pp. 299–324). Springer International Publishing. https://doi.org/10.1007/978-3-319-21094-0_18

Uher, J. (2016b). What is behaviour? And (when) is language behaviour? A metatheoretical definition. *Journal for the Theory of Social Behaviour*, *46*(4), 475–501. https://doi.org/10.1111/jtsb.12104

Uher, J. (2018a). Quantitative data from rating scales: An epistemological and methodological enquiry. *Frontiers in Psychology*, *9*, 2599. https://doi.org/10.3389/fpsyg.2018.02599

Uher, J. (2018b). Taxonomic models of individual differences: A guide to transdisciplinary approaches. *Philosophical Transactions of the Royal Society B*, *373*(1744), 20170171. https://doi.org/10.1098/rstb.2017.0171

Uher, J. (2018c). The transdisciplinary philosophy-of-science paradigm for research on individuals: Foundations for the science of personality and individual differences. In V. Zeigler-Hill & T. K. Shackelford (Eds.), *The SAGE handbook of personality and individual differences: Volume I: The science of personality and individual differences* (pp. 84–109). SAGE. https://doi.org/10.4135/9781526451163.n4

Uher, J. (2019). Data generation methods across the empirical sciences: Differences in the study phenomena's accessibility and the processes of data encoding. *Quality and Quantity. International Journal of Methodology*, *53*(1), 221–246. https://doi.org/10.1007/s11135-018-0744-3

Uher, J. (2020a). Human uniqueness explored from the uniquely human perspective: Epistemological and methodological challenges. *Journal for the Theory of Social Behaviour*, *50*(1), 20–24. https://doi.org/10.1111/jtsb.12232

Uher, J. (2020b). Measurement in metrology, psychology and social sciences: Data generation traceability and numerical traceability as basic methodological principles applicable across sciences. *Quality & Quantity. International Journal of Methodology*, *54*(3), 975–1004. https://doi.org/10.1007/s11135-020-00970-2

Uher, J. (2021a). Problematic research practices in psychology: Misconceptions about data collection entail serious fallacies in data analysis. *Theory & Psychology*, *31*(3), 411–416. https://doi.org/10.1177/09593543211014963

Uher, J. (2021b). Psychology's status as a science: Peculiarities and intrinsic challenges. Moving beyond its current deadlock towards conceptual integration. *Integrative Psychological and Behavioral Science*, *55*(1), 212–224. https://doi.org/10.1007/s12124-020-09545-0

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

Uher, J. (2021d). Quantitative psychology under scrutiny: Measurement requires not result-dependent but traceable data generation. *Personality and Individual Differences*, *170*, 110205. https://doi.org/10.1016/j.paid.2020.110205

Uher, J. (2022a). Functions of units, scales and quantitative data: Fundamental differences in numerical traceability between sciences. *Quality and Quantity. International Journal of Methodology*, *56*(4), 2519–2548. https://doi.org/10.1007/s11135-021-01215-6

Uher, J. (2022b). Rating scales institutionalise a network of logical errors and conceptual problems in research practices: A rigorous analysis showing ways to tackle psychology's crises. *Frontiers in Psychology*, *13*, 1009893. https://doi.org/10.3389/fpsyg.2022.1009893

Uher, J. (2023). *Overcoming method-centrism in psychology: Beyond rating scales*. Springer Nature Switzerland.

Uher, J., Addessi, E., & Visalberghi, E. (2013). Contextualised behavioural measurements of personality differences obtained in behavioural tests and social observations in adult capuchin monkeys (Cebus apella). *Journal of Research in Personality*, *47*(4), 427–444. https://doi.org/10.1016/j.jrp.2013.01.013

Uher, J., Asendorpf, J. B., & Call, J. (2008). Personality in the behaviour of great apes: Temporal stability, cross-situational consistency and coherence in response. *Animal Behaviour*, *75*(1), 99–112. https://doi.org/10.1016/j.anbehav.2007.04.018

Uher, J., & Visalberghi, E. (2016). Observations versus assessments of personality: A five-method multi-species study reveals numerous biases in ratings and methodological limitations of standardised assessments. *Journal of Research in Personality*, *61*, 61–79. https://doi.org/10.1016/j.jrp.2016.02.003

Uher, J., Werner, C. S., & Gosselt, K. (2013). From observations of individual behaviour to social representations of personality: Developmental pathways, attribution biases, and limitations of questionnaire methods. *Journal of Research in Personality*, *47*(5), 647–667. https://doi.org/10.1016/j.jrp.2013.03.006

Valsiner, J. (2001). Process structure of semiotic mediation in human development. *Human Development*, *44*(2–3), 84–97. https://doi.org/10.1159/000057048

Valsiner, J. (2012). *A guided science: History of psychology in the mirror of its making*. Transaction Publishers.

Valsiner, J. (2014). Needed for cultural psychology: Methodology in a new key. *Culture & Psychology*, *20*(1), 3–30. https://doi.org/10.1177/1354067X13515941

Valsiner, J. (2017). *From methodology to methods in human psychology*. Springer International Publishing. https://doi.org/10.1007/978-3-319-61064-1

Valsiner, J. (2019). *Social philosophy of science for the social sciences* (J. Valsiner, Ed.). https://doi.org/10.1007/978-3-030-33099-6

Valsiner, J. (2021). *General human psychology*. Springer Nature Switzerland. https://doi.org/10.1007/978-3-030-75851-6

van Geert, P. (2011). The contribution of complex dynamic systems to development. *Child Development Perspectives*, *5*(4), 273–278. https://doi.org/10.1111/j.1750-8606.2011.00197.x

Vautier, S., Lacot, É., & Veldhuis, M. (2014). Puzzle-solving in psychology: The neo-Galtonian vs. nomothetic research focuses. *New Ideas in Psychology*, *33*, 46–53. https://doi.org/10.1016/j.newideapsych.2013.10.002

Vessonen, E. (2017). Psychometrics versus representational theory of measurement. *Philosophy of the Social Sciences*, *47*(4–5), 330–350. https://doi.org/10.1177/0048393117705299

Vygotsky, L. S. (1962). *Thought and language*. MIT Press.

Wagoner, B., & Valsiner, J. (2005). Rating tasks in psychology: From a static ontology to a dialogical synthesis of meaning. In A. Gülerce, I. Hofmeister, G. Saunders, & J. Kaye (Eds.), *Contemporary theorizing in psychology: Global perspectives* (pp. 197–213). Captus.

Weber, M. (1949). In E. Shils & H. Finch (Eds.), *On the methodology of the social sciences*. Free Press.

Westerman, M. A. (2014). Examining arguments against quantitative research: 'Case studies' illustrating the challenge of finding a sound philosophical basis for a human sciences approach to psychology. *New Ideas in Psychology*, *32*, 42–58. https://doi.org/10.1016/J.NEWIDEAPSYCH.2013.08.002

Whitehead, A. N. (1929). *Process and reality*. Harper.

Wittgenstein, L. (1922). *Tractatus logico-philosophicus*. Kagan Paul.

Wood, A. M., Brown, G. D. A., Maltby, J., & Watkinson, P. (2012). How are personality judgments made? A cognitive model of reference group effects, personality scale responses, and behavioral reactions. *Journal of Personality*, *80*(5), 1275–1311. https://doi.org/10.1111/j.1467-6494.2012.00763.x

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, *45*, e1. https://doi.org/10.1017/S0140525X20001685

Young, B., Boccaccini, M. T., Conroy, M. A., & Lawson, K. (2007). Four practical and conceptual assessment issues that evaluators should address in capital case mental retardation evaluations. *Professional Psychology: Research and Practice*, *38*(2), 169–178. https://doi.org/10.1037/0735-7028.38.2.169

Zagaria, A., Ando, A., & Zennaro, A. (2020). Psychology: A giant with feet of clay. *Integrative Psychological and Behavioral Science*, *54*(3), 521–562. https://doi.org/10.1007/s12124-020-09524-5

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2017). Making replication mainstream. *Behavioral and Brain Sciences*, *41*, 1–50. https://doi.org/10.1017/S0140525X17001972

How to cite this article: Uher, J. (2023). What's wrong with rating scales? Psychology's replication and confidence crisis cannot be solved without transparency in data generation. *Social and Personality Psychology Compass*, *17*(5), e12740. https://doi.org/10.1111/spc3.12740