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

Target article

Psychometrics is not measurement: Unraveling a fundamental misconception in quantitative psychology and the complex network of its underlying fallacies

Jana Uher *

[1] *School of Human Sciences, University of Greenwich*
[2] *London School of Economics*

* Correspondence:
University of Greenwich, School of Human Sciences
Old Royal Naval College, Park Row, London SE10 9LS, United Kingdom
Telephone: +44(0)20-8331 9654; E-mail: mail@janauher.com

This research was funded by the European Commission (EC Grant Agreement number 629430).

Abstract

Psychometrics has always been confronted with fundamental criticism, highlighting serious insufficiencies and fallacies. Many fallacies persist, however, because each critic explores only some fallacies while still building on others. This article scrutinizes the epistemological, metatheoretical and methodological foundations of psychometrics, revealing a complex network of numerous conceptual fallacies underlying its framework of theory and practice. At its core lies a key challenge for psychology: the necessity to distinguish the phenomena under study from the means used to explore them (e.g., concepts, methods, data). This distinction is intricate because concepts constitute psychical phenomena in themselves and many psychical phenomena are accessible only through language-based methods. The analyses show how insufficient consideration of this important distinction and common misconceptions about concepts and language (e.g., signifier-referent conflation; reification of constructs) led to conflations of disparate notions of key terms in psychological measurement (e.g., 'variables', 'attributes', 'causality') and numerous interrelated fallacies (e.g., construct-referent conflation, phenomenon-quality-quantity conflation, numeral-number conflation). These fallacies are maintained and masked by repeated conceptual back-and-forth switching between two incompatible epistemological frameworks, *1)* an operationist framework of data modelling implemented through methodical and statistical operations and *2)* a realist framework of measurement sporadically invoked in theoretical considerations but neither theoretically elaborated nor empirically implemented. The analyses demonstrate that psychometrics constitutes only data modelling but not data generation or even measurement as often assumed and that analogies to (indirect or fundamental) physical measurement are mistaken. They provide theoretical support for the increasing criticism of psychometrics and its use in research and applied contexts.

Keywords
Psychometrics; Replicability; Latent variable; Psychological measurement; Quantitative method

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

Introduction

Developing methods to quantify properties of study phenomena is a central task in many sciences. Psychometrics, the field concerned with 'measuring the mind' (Borsboom, 2005), is aimed at making this possible for psychical phenomena despite the obvious differences from physical phenomena and inapplicability of physical measuring instruments. Over the last century, psychometrics has become a flourishing field with substantial commercial impact. But from the start, numerous lines of critique were voiced, highlighting serious insufficiencies and problems, such as the implicit but untested assumption of quantitative properties in psychical phenomena (Michell, 2008), the focus on correlational rather than causal relations for validation (Borsboom, 2005), the lack of representation theorems (Kyngdon, 2008a) and insufficient conceptual understanding of concepts and language (Maraun & Gabriel, 2013). Various fallacies were highlighted, such as erroneous equations of constructs with their referents (Slaney & Garcia, 2015) and of latent variables with constructs (Maraun & Halpin, 2008) as well as erroneous analogies with physical measurement ((Trendler, 2019; Uher, 2020a).

Although clearly recognized, these fallacies still persist in pertinent publications—even in the writings of scholars who are critically discussing their detrimental impact on psychological research. This is because all these fallacies are tightly interrelated and build upon each other, forming a complex network that underlies the psychometric framework of conceptual thinking and empirical practice. But each single critic typically focusses on just a subset of the fallacies already known, while implicitly still building on other fallacies in their thinking. This fragmentation entails that the entirety of fallacies and their complex interplay are not yet fully understood and that their effects on the theories and practices established in psychometrics have not yet been fully analyzed.

This article aims to put together the puzzle of the different lines of criticism voiced against psychometrics as measurement, to complement them with further fallacies still not well considered, and to highlight their interdependencies. This network of fallacies is central to the framework of psychometric theory and practice—and first establishes its functionality. Understanding its complexity and functioning is crucial for developing alternative approaches and methods of measurement in psychology that help avoid these fallacies in the future.

Transdisciplinary philosophy-of-science analyses

Psychologists, especially those concerned with measurement, often identify explicitly not as philosophers (Alexandrova & Haybron, 2016). But this is not a strength. It is a weakness. Psychology emerged from philosophy and—as every science—it builds on a philosophy of science, whether or not explicitly elaborated. The corresponding German term 'Wissenschaftstheorie' (theory of science) may describe this field more appropriately. Clearly both philosophy and theory are needed to scientifically explore the making of science, as this journal's name shows. Philosophical-theoretical research is essential for analyzing and overcoming the problems of psychological measurement (Haig & Borsboom, 2008). But many quantitative psychologists are still fairly reluctant to study the philosophy-of-science foundations of their own research and relevant concepts are often missing in publications.

Exploring the network of conceptual fallacies underlying psychometrics requires not only a philosophy-of-science approach but one that also cuts across disciplines in order to capitalize on and learn from other sciences' theories and practices of measurement, especially from those of metrology, the science of measurement, which is foundational for the physical sciences. But why should metrologists' concepts matter given that their physical study objects differ substantially from those of psychologists? Necessarily, different phenomena and properties require different approaches and methods (Bohr, 1937; Heisenberg, 1927; Uher, 2019); thus,

metrology need not matter at all for psychology. The point here however is that, when numerical data are generated and mathematically and statistically analyzed in order to describe and explore real-world phenomena, the generation of these data must be based on some transparent principles that are applied consistently across sciences and that ensure that the mathematical and statistical results obtained allow to make justified inferences about the study phenomena. This is essential for developing knowledge about these phenomena that can be set in relation to findings from other investigations about the same or other real-world phenomena—no matter in which science they may have been produced—thus, for establishing a secured knowledge base. It is also a matter of scientificity; a process as foundational to science as measurement cannot have entirely different meanings in different fields. Transdisciplinary analyses are therefore helpful to explore differences and commonalities in the measurement practices of different sciences and to identify basic principles that are applicable to all.

The present analyses rely on the *Transdisciplinary Philosophy-of-Science Paradigm for Research on Individuals* (*TPS-Paradigm*; (Uher, 2015c, 2018c) in which established concepts from various disciplines, complemented by novel ones, have been integrated into philosophical, metatheoretical and methodological frameworks that coherently build upon each other. These frameworks highlight connections, differences and communalities across sciences, and thus starting points for cross-scientific collaboration (Uher, 2020b). Together with its focus on research on individuals, including critical considerations of scientists' own role in research processes, the TPS-Paradigm therefore provides useful conceptual foundations for the present analyses. Some relevant concepts will be introduced below where needed; more information and references are provided in the footnotes, extensive elaborations elsewhere[1].

Necessarily, such transdisciplinary and philosophy-of-science frameworks require a terminology that is more abstract than that used for the specific theories, concepts, methods and approaches in the given disciplines and that inevitably diverges from any mono-disciplinary standard. This may feel unfamiliar to some readers. But for the present analyses this it is essential because key terms of psychological measurement codify conceptual fallacies and thus contribute to these fallacies' persistence in the field (Uher, 2021).

Critical realism: A philosophy for exploring the natural, social and experiential world

Many psychologists argue that, because measurement is about gaining knowledge about the natural world, psychological measurement must be grounded in scientific realism (Borsboom, 2005; Maul, 2013; Michell, 1999). The core idea of this epistemology is that both observable and unobservable phenomena can be explored and that true knowledge can be generated about them. This involves, *metaphysically*, the belief in the spatio-temporal existence of the world independent of its perception and conception by some conscious beings (mind-independent reality); *semantically*, the literal interpretation of scientific claims as describing this mind-independent reality; and *epistemologically*, the idea that thus-interpreted scientific claims yield true knowledge about that real world (Chakravartty, 2017).

[1] The TPS-Paradigm has already been applied *1)* to integrate and expand on previous concepts of individuals' psyche, behavior, language and contexts (Uher, 2013, 2015a, 2015c, 2016b, 2016a); *2)* to refine and newly develop concepts and methodologies for taxonomising and comparing individual differences in various kinds of phenomena and populations (Uher, 2015b, 2015d, 2015e, 2018b, 2018c), and *3)* to critically analyze concepts, theories and practices of data generation and measurement across the sciences (Uher, 2019, 2020a) and in quantitative psychology (Uher, 2018a, 2021). Applications are demonstrated in multi-method studies with humans and other species (e.g., (Uher, 2015b, 2018a; Uher, Addessi, et al., 2013; Uher & Visalberghi, 2016; Uher, Werner, & Gosselt, 2013). http://researchonindividuals.org.

In psychological measurement, however, scientific realism often takes the form of naïve realism involving assumptions that invariant quantities would exist in the world and independently of the methods used—ideas that even metrologists reject (Mari et al., 2017). More critical stances are needed, in particular, given that psychologists aim to explore individuals and the specific reality of their minds. How can this reality be mind-independent (Stent, 1969)? The point however is not, as some proponents of scientific realism seem to believe, to deny the existence of a reality, which includes the minds of the beings that form part of it; instead, it is about the ways in which we can explore that reality. The philosophical framework of the TPS-Paradigm comprises the presumption that all science is done by humans and that we can gain access to this reality only through our human perceptual and conceptual abilities (interpretations; Peirce, 1958, CP 2.308; Wundt, 1907), which inevitably limits our possibilities to explore and understand this reality. This should not be mistaken for ideas of radical constructivism (von Glasersfeld, 1991), which involve the assumption that knowledge could be developed without reference to a mind-independent ontological reality to which humans, as a species, have adapted over millions of years (Uher, 2015a). Instead, the presupposition made in the TPS-Paradigm highlights the fact that psychologists are always individuals themselves and thus, cannot be independent of their objects of research, unlike most natural scientists. Indeed, psychologists aim to explore minds—being equipped with nothing but a mind (Stent, 1969). This entails that psychologists' presuppositions about their study phenomena are (inevitably) influenced by the (explicit and implicit) beliefs that they have developed about these phenomena from their own, inherently anthropo-centric, ethno-centric and ego-centric experiences (Fahrenberg, 2013; Uher, 2015c, 2020b).

This epistemological stance comes close to that of *critical realism*, which emphasizes the reality of the objects of research and their knowability but also that our knowledge about this reality is created on the basis of our practical engagement with and collective appraisal of that reality (Bhaskar & Danermark, 2006). Hence, assumptions about the existence of reality must be distinguished from claims of the existence of truth because, without sentences, there is no truth. Sentences are elements of human languages, and human languages are human creations. "Truth cannot be out there—cannot exist independently of the human mind—because sentences cannot so exist or be out there. The world is out there, but descriptions of the world are not. Only descriptions of the world can be true or false. The world on its own—unaided by the describing activities of human beings—cannot" (Rorty, 1995, p. 5). What scientific communities establish as truth is thus a result of their consensus, designed from language games (Wittgenstein, 2009). That is, knowledge generation is not an increasing understanding of reality 'as it really is'. Instead, knowledge is increasingly useful to meet socio-practical demands—and therefore always theory-laden, socially embedded and historically contingent (Maraun, 1998; Maraun et al., 2009; Pinheiro, 2020).

Outline of the article

The article first discusses psychology's key challenge of distinguishing the phenomena under study from the means used for their exploration (e.g., concepts, methods, terms, data). To elaborate the intricacies of this distinction, it introduces metatheoretical concepts from the TPS-Paradigm that clarify the ontology of constructs (as one of psychology's most frequent study phenomena) as well as that of sign systems, in particular of language and data, and elucidate their functions for abstract thinking and science. These concepts are then used to elaborate in detail how insufficient consideration of this important distinction as well as common misconceptions about concepts and language led to conflations of disparate notions of key terms in psychological measurement and numerous interrelated fallacies. Thereafter, these

conceptual and terminological fallacies are used to scrutinize the concepts and practices of 'measurement' established in psychometrics. For this purpose, and to highlight commonalities and differences with metrology, the article introduces basic methodological principles from the TPS-Paradigm that were shown to underlie metrologists' concepts of measurement and measuring instruments and to be also applicable to psychological study phenomena. These principles are applied to demonstrate the ways in which the dense network of fallacies highlighted in this article underlies the pathways of reasoning in psychometrics that led to erroneous analogies with physical measurement and to the widespread but erroneous belief that psychometric modelling approaches could constitute measurement. The article closes by highlighting some far-reaching consequences and general directions for future developments.

Psychology's core challenges arising from their study phenomena's peculiarities

Psychologists face intricate challenges that arise from the peculiarities of their study phenomena (Uher, 2016a, 2020b). Unlike physical phenomena, for example, psychical phenomena are subject matter of a comprehensive everyday knowledge and language. But psychologists seldom reflect on the ways in which their lay psychology influences their own scientific psychology, such as when using everyday terms in language-based methods like rating scales (Uher, 2013, 2015d). Further challenges arise because psychologists' study phenomena involve also those by which all science is made (e.g., thinking, conceptualizing; Valsiner, 2012). All this requires psychologists to distinguish their study phenomena from the means (e.g., concepts, terms, methods, data) used to explore them—a distinction reflected in the terms psych*ical* versus psych*ological*[2] in the TPS-Paradigm (and many non-English languages). In psychology, however, this important distinction is seldom considered; indeed, it underlies many key fallacies that are inherent to psychometrics, as shown in this article.

To provide the necessary foundations for these analyses, this section's remainder briefly outlines metatheoretical concepts from the TPS-Paradigm that clarify the ontology of constructs and their function for exploring psychical and behavioral phenomena as well as the nature of sign systems like language and data and their roles for abstract thinking and science. These concepts will be applied, in the subsequent section, to explore conceptual problems and fallacies encoded in key terms of psychological measurement and to trace their possible origins in common misconceptions about concepts and language.

Constructs in psychology: Ontology and relevance

Ongoing psychical[3] phenomena, as well as many behaviors[4], are transient in nature and therefore conceived as occurrents (*perdurants* in formal ontology)—as *processes*. Of processual entities, only a part exists at a given moment; hence, they cannot be determined without knowledge of their previous occurrences. Occurrents are distinguished from continuants (*endurants* in formal ontology), which do exist in their entirety at any moment (e.g., material objects). As processes, psychical phenomena can be conceived only through *abstraction* from

[2] From Greek -λογία, -logia for body of knowledge (Lewin, 1936; Uher, 2016a). Analogously, we may get vir*al* (but not viro*logical*) infections and we do vir*ological* research.
[3] The *psyche* is defined in the TPS-Paradigm as the "entirety of the phenomena of the immediate experiential reality both conscious and non-conscious of living organisms" (Uher, 2015c, p. 431), with immediacy indicating absence of phenomena mediating their perception (Wundt, 1896).
[4] In the TPS-Paradigm, *behaviors* are metatheoretically defined as the "external changes or activities of living organisms that are functionally mediated by other external phenomena in the present moment" (Uher, 2016b, p. 490). This definition highlights essential differences in the accessibility of behaviors and psychical phenomena.

their occurrences over time, leading to concepts, beliefs and knowledge *about them*, which are psychical phenomena in themselves as well but different from and necessarily more stable than those they are about[5] (Uher, 2016a; Whitehead, 1929).

This explains why abstractions and complex ideas that are theoretically constructed by humans, called *constructs*, are among psychology's most frequent study phenomena (Maraun et al., 2009; Slaney, 2017). Their abstract theoretical nature entails that these conceptual entities often have several *construct referents*. Referents can be considered on different levels of abstraction; they may involve various concrete phenomena that are perceivable at a given moment (e.g., behaviors, emotions) but also conceptual entities (e.g., sub-constructs) that are each linked with their own set of more concrete referents. That is, referents can have nested conceptual structures in which meanings and referents can be 'inherited' from other concepts (Uher, 2021). Abstraction involves that construing persons emphasize some aspects of the referents they consider, while deemphasizing others. Therefore, any given construct cannot reflect its referents in the same ways as these can be perceived at any moment (Vygotsky, 1962; Whitehead, 1929). Differences in the particular referents, aspects and levels of abstraction that persons (implicitly) consider enable unparalleled proliferation and complexity—and thus changeability and diversity in the constructs created.

This highlights another key challenge in psychology that is still not well considered. The *scientific concepts* used to explore psychical phenomena *constitute psychical phenomena in themselves*—and thus do not exist outside of the empirical systems under study (Uher, 2020b). This substantially complicates the important distinction between the study phenomena and the means used for their exploration. Further complications arise from human language, which is inevitable for doing science and many psychological study phenomena are accessible only through language.

Language and its essential function for abstract conceptual thinking

Language plays a key role in science because scientists must describe and explain their study phenomena. Moreover, language and other sign systems have important functions for abstract thinking. They enable us to represent perceivable phenomena and their properties (e.g., two large owls) in single words (e.g., 'two', 'large', 'owls'). In words, we can make concrete entities independent of their immediate perception and abstract them into objects of consideration (conceptual entities)—thus, *reifying* them (e.g., 'size'; 'birds'). Through this *hypostatic abstraction* (Peirce, 1958), CP 4.227), we develop words that refer to concrete referents also in their absence, thus abstracted from the here and now. We can also develop words that have abstract referents, such as concepts and ideas describing phenomena and properties distant from immediate perception (e.g., 'animals') or imperceptible in themselves (e.g., 'quantity'; (Uher, 2015a, 2016b). That is, every word is a concept in itself (Vygotsky, 1962). This contributes to the intricacies of distinguishing the phenomena under study from the means needed for their exploration.

These challenges also concern the data systems that scientists generate about their study phenomena. So, what actually are 'data' and what function do they fulfil in science?

[5] Transient psychical events (e.g., thoughts, emotions), called *experiencings* (Erleben) in the TPS-Paradigm, can be distinguished from temporally more persistent phenomena (e.g., beliefs, mental abilities), called memorized psychical resultants or *experiences* (Erfahrung; with memorization referring to any retention process). But these latter can be accessed only in individuals' experiencings (Uher, 2016a), and must be reconstructed in each moment anew within the given context, whereby they are adapted and changed before becoming memorized again (Schacter & Addis, 2007).

Data and their function in science

Data[6] are signs (symbols, mathematical or semiotic representations; e.g., Indo-Arabic numerals, Latin or Greek letters) that scientists use to indicate information about the study phenomena and properties and to which they attribute particular meanings (e.g., numbers, qualities). As signs (e.g., variable names and values), the function of data is to represent in physically persistent ways (e.g., printed or digital spreadsheets) information about properties and phenomena as conceived by the data-generating persons. This representational function of signs is so deeply engrained in our language and thinking that we seldom become aware that any sign—every word of natural language, every scientific term, and every mathematical symbol—comprises three distinct components (Figure 1a). These are *(1)* a physical, publicly accessible component (e.g., phonetic or visual patterns like /faɪv/, *5, V,* IIIII) used as *signifier* that symbolically represents *(2)* the *referent*, the actual object of consideration to which it refers (e.g., phenomenon, property, construct), and *(3)* the *meaning* (*signified*) that both have for the sign-using persons in a given context (Salvatore, 2019), which is a psychical phenomenon in itself (see similarly, Ogden & Richards, 1923). We become acutely aware of these three components when learning a foreign language because this involves learning the specific signifiers used in that language (e.g., Eule, chouette, gufo, нея́сыть[7]), their relations to particular referents (e.g., owls) and the meanings that they may have for speakers of that language (e.g., a bird, colloquial for a night person). It is the tight conceptual relationships among these three components that first establish the functionality of this triadic composite as a sign (Figure 1a).

Conceptual fallacies and disparate notions of key terms

The just-introduced metatheoretical concepts of constructs, psyche, behavior and of sign systems like language and data are applied in this section to explore common misconceptions of concepts and language and to highlight their role in the conceptual fallacies inherent to psychometrics. The analyses will also reveal that these fallacies are codified in and thus maintained by important key terms in the field.

Signifier–sign equation involves signifier–meaning and signifier–referent conflations

In everyday life, but also in science, the tight triadic interrelations among the three components of signs (see above) are seldom explicitly considered. This entails various fallacies such that signs (e.g., words, data, terms) are often—and be it just implicitly—equated with their most directly apparent component, their signifier. This *signifier–sign equation* ignores the signifier's intricate relations to a referent as well as to the meaning linking both and without which a signifier (literally) could not signify something (Figure 1b). This fallacy often occurs in language-based methods, such as those used in psychometrics in which verbal information is presented to respondents who must respond in likewise verbal form (e.g., rating scales, IQ tests). It is reflected, for example, in the belief that "what really matters in validity is how the test works, and this is […] a property […] of the *measurement instrument itself (i.e., of the concrete, physical thing that you can drop on your feet, rather than of a linguistic entity*)" [8] (italics added; Borsboom et al., 2009, p. 149).

[6] The term *data* is used inconsistently in psychology; see the section Disparate notions of 'variables'… below.
[7] These signifiers mean owl in German, French, Italian and Russian.
[8] This statement implies that digital and paper versions of the same scale, given their different physical properties, could also differ in their measurement properties—a, for the digital age, surprising statement.

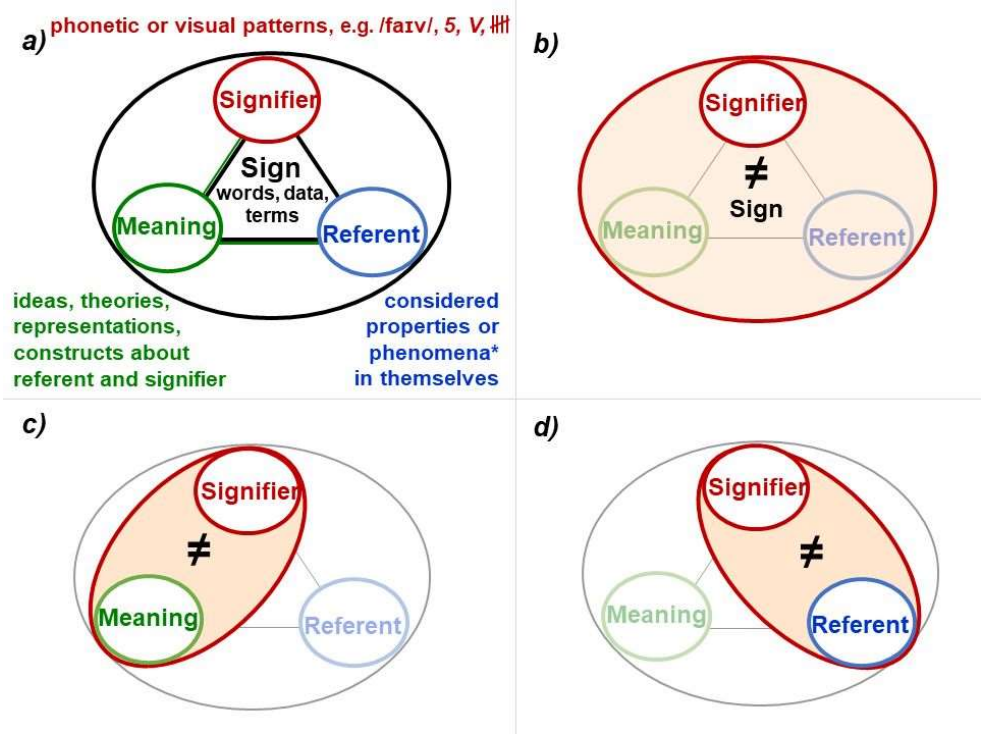

Figure 1. Sign systems, their three components, and common fallacies about them. (a) The three components of sign systems; (b) signifier–sign equation; (c) signifier–meaning conflation; and (d) signifier–referent conflation. This includes conceptual phenomena (e.g., other constructs) as well.

Signifier–sign equation involves that people also erroneously assume that a sign's meaning would be contained in the signifier itself. This *signifier–meaning conflation* (Figure 1c) is reflected, for example, in the widespread assumption that standardizing signifiers (e.g., writing down item wordings) could allow to standardize also their meanings across persons, times and contexts. But this ignores pronounced individual variation in item interpretations—thus, subjectivity—that occurs even if all psychometric criteria are met (Lundmann & Villadsen, 2016; Rosenbaum & Valsiner, 2011; Uher, 2018a; Uher & Visalberghi, 2016).

Signifier–sign equation furthermore involves that the signifier is mistaken for its referent, thus *signifier–referent conflation* (Figure 1d). But signifiers do not carry their meanings in themselves; signifiers are largely arbitrary and can therefore denote different referents (e.g., *M, D, C, L* can signify letters or numerals). The triadic concept of signs (Figure 1a) thus highlights that the meaning of particular data is not given by their signifiers. Instead, signifiers are interpreted as representing particular phenomena and properties (referents) *only* given particular theories and expectations (meaning[9]). With different theories, data may be interpreted differently (Van Fraassen, 2012); this explains why data are always theory-laden (Boon, 2015; Kuhn, 1962).

These conflations of the three distinct components of sign systems are widespread in everyday life. They led to many conceptual fallacies in psychometrics as elaborated now.

[9] To highlight the meaning component's essential role for establishing the triadic interrelations, sign systems are called semiotic *representations* in the TPS-Paradigm (Uher, 2015c, 2015a).

Disparate notions of the term 'variable' entail variable–referent conflation

The term 'variable' is central to psychological measurement. It has, however, two disparate notions that are seldom clearly distinguished and that are, due to signifier–referent conflation, often conflated with one another. This conflation lies at the bottom of major fallacies in psychometrics. Specifically, 'variables' denote variability. Variability in phenomena and properties is often in the focus of research because it can reveal important information about their structures and functionings. Likely for this reason, psychologists often refer to their *study phenomena and properties in themselves* as 'variables' (e.g., age, 'extraversion', 'neuroticism', psychical phenomena located in people's heads). This may explain why many psychologists speak of studying 'variables' rather than of studying phenomena or properties (see Maraun & Gabriel, 2013). But, at the same time, psychologists also speak of 'variables' in terms of the elements of their data sets that they subject to statistical analysis (e.g., item variables, data sets in variables by cases format; variable-oriented analyses), thus as the *semiotic encodings of their study phenomena and properties* (e.g., located on computers). In metrology as well, variables constitute the results of measurement but not the properties to be measured (Mari et al. 2015; Uher, 2020).

But importantly, these two notions are not just different—they are mutually exclusive. The former refers to the phenomena and properties under study in themselves, the latter refers to their semiotic encodings. Whatever notion one may prefer, 'variables' cannot be *both* the empirical system under study *and* the symbolic (data) system analyzed in lieu of the former. This is a prime example of the challenges that language-based methods entail for the essential distinction of the study phenomena from the means used for their exploration. Their failed distinction entails conceptual fallacies in which scientists frequently (without necessarily being aware of this) switch conceptually between these two disparate notions of 'variables'. This conceptual switching, promoted by scientific (or even naïve) realist beliefs, leads scientists to conflate the phenomena and properties under study with their semiotic encodings, and thus entails *variable–referent conflation* (with 'variables[10]' denoting semiotic encodings; Figure 2; (Maraun & Halpin, 2008; Uher, 2015e; Woodward, 2011).

Here, the term (data) variables is used in the sense of semiotic encodings, whereby the inserted "(data)" is intended to remind readers of this notion. However, the term '*data'* itself is afflicted with disparate notions and signifier–referent conflation. It often denotes semiotic encodings (mathematical representations, e.g., variable names, values) as used here. But it also seems be used for the sample itself, thus the specific set of phenomena, properties and individuals studied from the universe of all possible ones (e.g., in Borsboom & Mellenbergh, 2004)—in line with the meaning of 'variable' as denoting the research objects in themselves.

Construct–referent conflation

A further fallacy that is promoted by sign-signifier equation and signifier–meaning and signifier–referent conflations of words in general and that is therefore widespread in research with language-based methods such as psychometrics is to conflate constructs as *theoretical-logical-linguistic thinking tools* with *their construct referents*, thus with the real-world entities that they are meant to denote (Danziger, 1997). Here again, the study phenomena are not distinguished from the concepts and methods used for their exploration. This *construct–referent*

[10] Ideally, the term 'variable' should be replaced by two others to clearly distinguish its two distinct notions.

conflation[11] (Figure 2; Slaney & Garcia, 2015) occurs, for example, when scientists interpret constructs as reflecting 'attributes' or qualities that individuals 'possess' (e.g., in Cronbach & Meehl, 1955), thus ascribing them an ontological status, as widely done with 'trait' constructs (Uher, 2013, 2015e). The construal of constructs allowed scientists to turn abstract ideas into entities, thereby making them conceptually accessible to empirical study. But this *entification* misguided psychologists to overlook their constructed nature (Slaney & Garcia, 2015) and to focus primarily on methodological and methodical approaches, ignoring the necessity to develop their epistemological foundations as well (Hanfstingl, 2019).

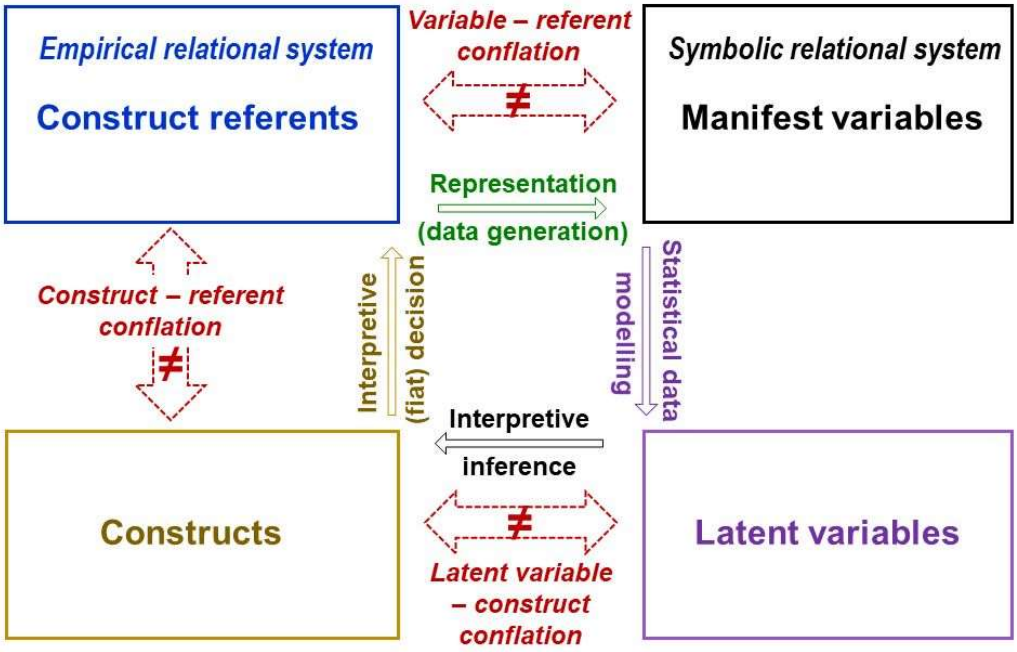

Figure 2. Conceptual fallacies derived from misconceptions of language and concepts. The term variable here means semiotic encodings.

Latent variable–construct conflation

We can create words to denote entities construed on all levels of abstraction—and thus also labels for the (data) variables used to semiotically encode such conceptual entities. (Data) variables encoding abstract conceptual entities are often conceived as *latent (data) variables*, such as the synthesized variable 'extraversion' summarizing scores of several *manifest (data) variables* that each encode more specific entities (e.g., behaviors). The availability of single-word labels for abstract conceptual entities (e.g., 'extraversion', 'neuroticism'), however, often misleads people to treat them as concrete entities that are located in the world or inside people (fallacy of misplaced concreteness; Whitehead, 1929; also Peirce, 1958). This reification of linguistic abstractions is promoted through the disparate yet often not distinguished notions of

[11] Slaney and Garcia termed this construct-*entity* conflation; but an entity can also be a construct (conceptual entity) in itself and construct referents can be other constructs as well.

the term 'variable'. In conjunction with variable–referent conflation, this leads to erroneous conflations of (data) variables denoting constructs (called 'collective variables'; Thelen & Smith, 1994) with these constructs in themselves, thus to *latent variable–construct conflation* (Figure 2; Maraun & Halpin, 2008). A (data) variable labelled 'extraversion' is then conflated with the abstract concept of 'extraversion' that it encodes and this, in turn, is conflated with its various referents in participants' experience and behavior (construct–referent conflation*;* Figure 2).

Kinds of referents encoded in (data) variables: Implications for statistical analysis

For meaningful interpretations of statistical analyses of (data) variables (located on computer) *with regard to* the real-world phenomena and properties under study (e.g., located in individuals), it is essential to know what exactly is encoded in the (data) variables. This is because their statistical relationships are being analyzed *in lieu of* those of the phenomena and properties under study in order to unravel possible causal relations among those latter. The fact that psychologists encode in (data) variables all kinds of referents and on all levels of abstraction entails problems that seriously undermine the meaningfulness of statistical analyses (on computer) and their interpretation *with regard to* the actual phenomena and properties under study (in individuals) and their interrelations (Toomela, 2008).

Specifically, constructs have diverse referents each featuring various aspects that are considered and emphasized differently; therefore, constructs cannot represent any of these referents as they may occur in individuals. The encoding of constructs in collective variables entails that possible causal relations of those construct referents (*explanantia*), individually or collectively, with the phenomenon to be explained (*explandum*) cannot be explored. This is because, in collective variables, it is impossible to differentiate the relation of a single referent from the whole construct system in which it may be involved. This may mask the actual causal relations so that statisticians may misleadingly understand the composite of referents as a cause instead of just one single referent, or vice versa, one single referent instead of the whole composite. Causality may also be erroneously attributed to one particular referent although another referent of the composite is actually important. For example, the widely used construct of sex/gender may refer, amongst others, to genetic, hormonal, bodily, behavioral and cultural differences, but which ones are actually relevant and causally related to a given target phenomenon and in what ways cannot be analyzed with this collective (construct) variable (for details and further problems, see Toomela, 2008)

These problems are of particular concern for psychometricians because they focus on constructs to study psychical and behavioral phenomena and because many of the rating item variables therefore used describe not concrete but conceptual entities (e.g., 'habitual behaviors'; (Uher, 2020a). This concerns all psychometric methods of analysis that are aimed at exploring causal relationships, thus latent variable models (Borsboom, 2008), latent network models and their combinations (Epskamp et al. 2017; Guyon et al., 2017) alike.

Disparate notions of the term 'attribute' and their conflation

The problems entailed by the undifferentiated use of term 'variable' are fortified by analogous problems with the term 'attribute', another key term with disparate notions that are frequently conflated with one another leading to conceptual fallacies. Specifically, it is variously stated that 'psychological attributes' "can be said to be real" and to "exist as emergent features of conscious beings", such as being in pain or hungry (Maul, 2013, p. 756) and that there are 'attributes of persons', such as extraversion, emotional intelligence, knowledge, working memory capacity, beliefs, intentional states and gender, as well as 'composite attributes' depicted as concepts or terms that refer to sets of such 'person attributes' (Maul, 2013). 'Attributes' are

assumed to be quantitatively structured (Michell, 1999) and to causally produce variations in the results obtained from measurement procedures (Borsboom, 2005). It is stated that there can be within-person and between-person variation in 'attributes' and that individuals can have a position on an 'attribute' (Borsboom & Mellenbergh, 2007). The term 'attributes' is also used to denote properties of physical objects (e.g., temperature, velocity; Maul, 2013). These are just some examples of the disparate notions ascribed to the term 'attribute', denoting either *1)* psychical phenomena in themselves; *2)* constructs and terms about them; *3)* properties that may occur in study phenomena and objects, and that may be quantitative and may interact with measuring instruments; as well as *4)* (data) variables that semiotically encode *1)* to *3)*.

The conflation of these disparate notions is yet another example of the failed distinction of the phenomena and properties under study from the means used for their exploration. This example also illustrates the codification of this failed distinction, and thus its maintenance, in a key term of the field.

Phenomenon–quality–quantity conflation

The disparate notions of key terms are often masked by the common nominalization of 'variables', 'attributes' and constructs (e.g., 'traits'; Slaney & Garcia, 2015), and the entification that this entails. Entification also masks another key fallacy in psychological measurement—the frequent failure to specify the phenomena, qualities and quantities studied. Psychologists often ignore that phenomena (or objects) in themselves cannot be measured; only properties can be. Any phenomenon (or object) typically features various properties of different qualities. Behaviors, for example, have temporal and spatial properties; they may also be ascribed constructed qualities of social desirability, amongst others. Bricks have the properties of length, weight, density, hardness, color, and temperature, amongst others. Therefore, scientists must specify in their study phenomena (objects) the particular quality of interest; one cannot just measure a behavior or a brick.

This specification is also a precondition for measurement because quantities are always of something—a quality. *Qualities* (from Latin *qualis* for "of what sort, of such a kind") are properties differing in kind, whereas *quantities* (from Latin *quantus* for "how much, how many") are divisible properties of entities of the same kind—the same quality (Hartmann, 1964). Quantities are qualitatively homogenous; adding or dividing their magnitude does not change their meaning; for example, adding entities of length keeps their quality as being that of length unaltered, whereas this is not possible for perceived color. Divisible properties of the same quality differ only quantitatively, never qualitatively (Michell, 2012).

Entities of the same quality can be compared in their divisible properties (quantities) regarding that quality. In behaviors, divisible properties can be identified in their temporal and spatial qualities (Uher, 2015b, 2018a); for empirical examples, Uher, Addessi, et al., 2013). But in the manifold qualities of experiencing, what properties could be divisible? What divisible properties could there be in the abstractions needed to study these transient processual phenomena? And how could divisible properties be identified in constructs, thus in abstract conceptual entities that have heterogeneous real-world referents, each featuring different and also differently emphasized qualitative properties?

Confusion about heterogenous versus homogenous 'orders'

The fact that the qualitative heterogeneity of psychical phenomena precludes their measurement was recognized, amongst others, already by Binet end of the 19th century, as Michell (2012) highlighted. Both use the term heterogeneous 'order' for the hierarchical organisation of constructs by their level of generality and abstraction, such as in intelligence

research or in the biological classification of species, as Michell illustrates. But contrasting this taxonomic 'order' with homogeneous 'order' in terms of positional information about a series of magnitudes of the same quality (Michell, 2012) implies that these disparate notions of 'order' would be somehow comparable. In terms of taxonomic order, the *same* person can be said to be a human, primate, mammal, vertebrate and animal; all this concerns only more abstract levels of consideration of the *same* entity. Ordinality, by contrast, refers to relations among *different* entities, such as *different* persons' body length, which can be ordered according to their magnitude. The undifferentiated use of the term 'order' for both taxonomic classification and ordinality obscures essential conceptual differences, thereby contributing to the common belief that constructs could be measurable and thus thwarting Michell's long-standing efforts to clarify misconceptions about the measurability of psychical phenomena.

Measurement: Most basic concepts, principles and processes

The conceptual fallacies and disparate notions of key terms of psychological measurement analyzed up to here are now used in this section to scrutinize the concepts and practices of 'measurement' established in psychometrics, highlighting commonalities and differences to concepts of measurement from metrology.

Measurement requires "one to spell out … why one is treating the data patterns as measurements" (Kyngdon, 2008b, p. 109)—that is, some basic criteria. Metrologists define measurement as a structured process, comprising operative structures to assign numerical values to the *measurands* (entities to be measured) in reliable, valid and explicitly justified ways (Mari et al., 2017). Similarly, psychologists demand "observational procedures that are set up to work in such a way that, if an object of study is subjected to them, they generate output that carries information about the state, value, position, or constitution of the system with respect to the concept[12] of interest" (Haig & Borsboom, 2008, p. 2). Measurement is thus an approach of data generation, targeted at obtaining quantitative information about study phenomena. As elaborated above, data (signs) are semiotic representations that can be stored, manipulated, decomposed and recomposed (e.g., on computer), that is, *analyzed in lieu of the actual properties and phenomena under study* (the referents) and in ways not applicable to these latter (in individuals). But inferences to these latter can be made *only if* the data represent relevant properties of the study phenomena in appropriate ways. This highlights the necessity of a basic representation theorem.

A basic representation theorem

Measurement is the assignment of numerical values such that the conceptual properties and interrelations of the (lexical and numerical) signs used as data (e.g., variables and their values on computer) appropriately represent the study phenomena's empirical properties and their interrelations (e.g., in individuals). This idea is basic to representational theory of measurement (RTM; Krantz et al., 1971). It formalizes axiomatic conditions by which empirical relational structures can be mapped to symbolic relational (data) structures (*representation theorem*), including permissible ways for transforming the latter without breaking their mapping onto the former (*uniqueness theorem*; Kyngdon, 2008a; Narens, 2002; Vessonen, 2017). But RTM stipulates neither theoretical concepts nor procedures for how and why any given empirical property could be mapped to a symbolic relational system, and many of its notions remain vague (Borsboom & Scholten, 2008; Mari et al., 2017).

[12] Note this instance of construct (concept)–referent conflation.

Still, and although clearly insufficient as a measurement theory, RTM stipulates the most basic principles underlying any kind of data generation in that information about study phenomena and their (qualitative and quantitative) properties (e.g., located in individuals) is encoded in signs (e.g., data variables on computer). This highlights that representation theorems also underlie the data generation methods used in psychometrics. Psychological assessment methods, especially rating scales, are however fraught with methodological, conceptual and terminological problems as explored now.

Disparate notions of 'behaviors' and 'responses' in psychological assessment

In psychology, given the challenges in distinguishing the study phenomena from the means used for exploring them, misconceptions frequently emerge about what actually constitutes the empirical relational system in a study. These misconceptions are promoted by the widespread labelling of both psychical *and* behavioral phenomena as 'responses' or ('overt' and 'covert') behaviors, likely misguided by the immaterial, transient and processual nature of both (Uher, 2016b). This undifferentiated terminology misleads researchers to ignore fundamental differences in these phenomena's accessibility that profoundly impact research methodologies (Uher, 2019). It also entails another fallacy in psychological assessment. Specifically, when psychologists label both individuals' finger movements for pressing buttons or ticking scales (Baumeister et al., 2007) *and* the mental processes involved in a given task as 'behaviors' or 'responses', this blurs the distinction what actually constitutes the empirical relational system and which part of the data generation process requires a representation theorem as demonstrated now by the example of rating scales.

Failed implementation of representation theorems in rating scales

In rating methods, neither raters' finger movements in themselves nor their ticks on the scales (as these behaviors' residuals) nor raters' judgement processes constitute the empirical relational system. These are just operational procedures involved in the mapping process in which raters match the outcomes of their judgements about the actual empirical system of interest (e.g., individuals' behaviors, feelings or beliefs) to the symbolic relational system provided (e.g., rating scales on sheet). That is, the complex task of executing a data generation process that is intended to meet measurement criteria is delegated to raters, thus to lay people commonly unfamiliar with the measurement theories therefore needed. Raters are provided with neither information about these theoretical foundations nor instructions of how these could be applied to psychical and behavioral phenomena. By contrast, in physical and behavioral (ethological) measurement, measurement-executing persons are instructed and trained about how to use the given operative procedures for generating results.

Ratings do not even constitute a method that—at least theoretically and with the necessary knowledge and instruction—could enable measurement at all. Indeed, rating methods fail to implement even a most basic representation theorem because rating items and scale categories serve *both* as descriptions of the empirical relational system (e.g., behavior and experience in individuals) *and* as symbolic relational system (e.g., item variables on sheet), leaving the two relational systems and the mapping relations between them unspecified (Uher, 2018a). These specifications are left to implicit decisions made by respondents, which remain largely unexplored. Psychometricians commonly make only general assumptions about raters' inattention, possible faking and response bias—unrelated to any specific referents to be judged, thus unrelated to the actual phenomena under study.

Instead, psychometricians focus on the permissible transformations of rating data as specified in Stevens' (1946) scale types (uniqueness theorems), likely because these stipulate at least some concrete quantitative concepts (later refined through latent variable theory, see below). Given Steven's (1946) simplified definition of measurement as the "assignment of

numerals to objects and events according to rules" (p. 677), researchers recode raters' choices of (lexically labelled) answer categories into numerals in always the same ways for all items, regardless of the phenomena and properties to which they refer. The uniqueness theorems, as specified through Steven's scale types, are then regarded as the theoretical justification for interpreting these numerals as numbers.

Numerals are signifiers ("a black mark on a piece of paper or certain sounds which I utter", Campbell, 1920, p. 267) and thus arbitrary (e.g., 1, 5; I, V). *Numbers*, by contrast, are mathematical objects arising from ontological interrelations among real phenomena (Hartmann, 1964). Numerals are often used to represent numbers but also just order (e.g., 1st, 5th) or only categorical—i.e., qualitatively different—properties that have no quantitative meaning at all (e.g., room 'numbers'; Campbell, 1919/2020). Recoding rating scale categories in order to create 'quantitative' data thus involves *numeral–number equation*, another instance of signifier–meaning conflation. In conjunction with the conflation of the different notions of 'response' and 'behavior', this leads psychologists to overlook that it is the raters who actually generate the data, whereas researchers' recoding of scale categories is just the transposition of one symbolic system into another (Kyngdon, 2008a)—thus, only a step of *data processing* but not one of data generation (Uher, 2018a, 2019, 2020a).

Correct responses, reaction times and their relations to psychical phenomena

Unlike ratings, many psychodiagnostic methods record responses that have a *fixed agreed meaning*, such as correctness (e.g., in intelligence, educational and achievement test, problem-solving tasks, attention and concentration tests). Psychologists also often record *reaction times*, thus physical responses (e.g., in implicit association tests, flanker tasks, go/no go tests, lexical decision tasks). These methods simplify the assignment task that respondents must accomplish. They allow scientists to record (i.e., semiotically encode) the occurrence of correct responses and response times, which however are only *outcomes* of the psychical phenomena involved in their emergence. Hence, in these methods, it is these responses that constitute the empirical relational system to which the symbolic relational system is mapped, but not the psychical phenomena in themselves, thus not the actual phenomena under study. These latter can *only be inferred* from the responses recorded. But externally similar responses may emerge from different internal phenomena and different processes among them (i.e., they are polygenetic). Expected responses are therefore no guarantee that specific psychical processes have taken place. Moreover, the whole psychical system cannot exist without various subprocesses that must be present for a given process to emerge at all (e.g., perception, long-term memory). That is, the existence of an internal phenomenon is no guarantee that it will also become manifest in observable outcomes and in expected ways. As a consequence, one-to-one inferences from recorded outcomes to the actual phenomena of interest cannot be made (for details, Toomela, 2008).

These profound challenges must be considered when establishing measurement processes targeted at psychical phenomena. Some psychologists suggested psychometric approaches would be analogous to those that metrologists have developed for measuring physical properties that are not directly accessible. The remainder of this section introduces basic methodological principles underlying metrologists' concepts of measurement and measuring instruments that are needed to scrutinize this assumption in the next section.

Two basic methodological principles of measurement

Metrologists elaborated a structural framework of physical measurement, involving epistemology, methodology and methods needed to devise processes that establish causal measurand–result relations (Mari et al., 2017). Previous analyses using the TPS-Paradigm

(Uher, 2020a) showed that, on a methodological level, this metrological framework builds on two basic principles—data generation traceability and numerical traceability.

Data generation traceability: Object-dependent measurement processes

The first methodological principle requires that the ways in which results are assigned to the quantity to be measured (measurand) in the study phenomena (objects) must be made fully transparent and thus traceable. To justify that the generated results are attributable to the measurands, measurement processes must be designed from knowledge about the study objects and their properties (called *object-dependence* or *object-relatedness* in metrology; Mari et al., 2017). This involves explanations how the specific operative structures allow to make numerical assignments such that they reveal reliable and valid information about these measurands, and only about them and not also on other influence properties (Mari et al., 2015). This knowledge must be implemented in *unbroken documented chains of comparisons* that connect the measurand with the result. At each step, the entities of the connected properties can be compared with one another regarding their quantities so that quantitative information from one property can be converted into quantitative information in another property, thus establishing *proportional relations* between the *quantities* of the different properties involved (see thermometer example below).

Numerical traceability: Subject-independent results linked to known standards

The second methodological principle of measurement underlying metrological frameworks requires that the numerical value assigned to the measurand is also linked to known standards, in likewise documented and transparent ways. The process design must ensure that results are invariant with respect to the persons (subjects; e.g., operators, users) involved (called *subject-independence* or *inter-subjectivity* in metrology). This means that results must be reliably interpretable and always represent the same information about the measurands across time and contexts (Mari et al., 2017). To ensure that the results have the same meaning everywhere (e.g., specific length of 1 meter), metrologists establish unbroken documented conversion (calibration) chains from primary references (e.g., international prototype meter) to all working references (e.g., meter rules) used for measurement in non-metrological research and everyday life (JCGM200:2012, 2012). Psychologists must develop (and have in parts already done so) analogous ways to establish an intersubjective meaning for their numerical results (e.g., time-based measurements of behavior; answer categories with universally agreed meanings of correctness in educational and achievement tests).

These two methodological principles are fundamental for measurement. But explicit analogous concepts had so far been lacking in psychology although both principles are—on their abstract methodological level—meaningfully applicable also in psychology (see below; (Uher, 2020a). Moreover, both principles are also essential for instrument development.

Measuring instruments: Conceptual foundations

Instruments for physical measurement

Technical instruments are devised to minimize the involvement of human abilities in measurement processes and to enable the measurement of physical properties that humans can perceive not directly (e.g., density) or not accurately enough (e.g., temperature). In such cases, operational procedures that systematically connect the measurand with the result are implemented via mediating properties, whereby the measurand empirically interacts with the first mediator in the chain, which in turn may interact with a further one, and so on. Thermometers, for example, structurally connect the properties ‘temperature’ and ‘spatial extension of tubed

mercury' though physical laws[13]. The latter is connected with 'length of extension over scale' through visual comparison, and this, in turn, is connected with (data) variables and values' through semiotic encoding[14] (Uher, 2020a).

What is the measuring instrument in psychological assessments?

Metrologists conceive the psychologists' generation of quantitative data *directly by persons* (e.g., raters, observers) as 'human-based measurement', 'humans as measurement instrument' (Pendrill, 2014), and 'persons as data generation systems' (Berglund et al., 2012). Thus, for metrologists, the person is the measuring instrument. Psychologists, by contrast, regard the items (statements, tasks) and answer categories as their 'measuring' instruments. Given this, they direct all efforts for instrument development and improvement at their tests and rating scales (e.g., psychometric properties) but not at the respondents' abilities and knowledge for executing the data generation process as metrologists would do given their understanding of the persons as constituting the instruments (Uher, 2019).

This conceptual difference illuminates further fallacies in psychometrics. Specifically, many psychologists assume that invariant quantities exist in their study phenomena (Michell, 2008) and independently of the methods used (Borsboom & Mellenbergh, 2004). These naïve realist views entail the belief that ideal methods (e.g., rating scales) could allow to empirically implement an identity function that turns pre-existing 'real' quantities into estimated (manifest) scores (with definable errors or probabilities). But interactions between study property and method always influence the results obtained (Bohr, 1937; Heisenberg, 1927; Mari et al, 2017). In psychological investigations, these interactions are intricate because they are mediated by the data-generating persons who interact with (e.g., perceive, interpret—the meaning) both the study phenomena and properties (e.g., behaviors, experience, frequencies—the referents) *and* the methods used (e.g., rating scales—semiotic encodings, here their signifiers). Both metrologists' and psychologists' notions of instruments fail to conceptualize these complex interactions, which derive from and thus reflect the triadic interrelations among signifier (symbolic system), referent (empirical system) and the meanings that both have for particular persons in particular contexts and which are essential for establishing these interrelations (and thus assignment relations between both systems). That is, establishing data generation and numerical traceability requires knowledge of raters' understanding and use of items scales, which vary substantially (Uher, 2018a), as well as careful consideration of the intricate challenges that language-based methods entail for the important distinction of the study phenomena from the means used for exploring them.

Psychometrics as measurement? A key assumption scrutinized

The metatheoretical and methodological concepts of measurement and measuring instruments across sciences are now applied in this final section to scrutinize some scientists' assumptions that psychometrics could constitute measurement. Specifically, they are used to scrutinize assumptions psychometrics would be analogous to metrological approaches of indirect or even direct physical measurement.

[13] Hence, instrument design requires knowledge of systematic (lawful) connections among properties, identifiable experimentally (Mari et al., 2017). But often, this knowledge is developed only during instrument development; thus, both processes are iterative and inform each other (Boon, 2015).

[14] These two latter steps are also often automatized to further reduce the involvement of human perceptual and conceptual abilities in measurement processes (Uher, 2020a).

Mistaken analogies of psychometrics with physical measurement

In *fundamental (direct) measurement*, metrologists determine physical quantities through direct comparison and evidence of concatenation (e.g., length of rods). For those physical properties for which this is not possible because they are not directly accessible, metrologists developed various approaches of *indirect measurement*. Specifically, in *derived measurement*, properties are measured indirectly from relations between other properties, which must be measurable in themselves either directly or independently (e.g., density from mass and volume; Campbell, 1920). In *associative measurement*, scientists assume that the quantity to be measured (e.g., temperature) is systematically connected (associated) with another, independently measurable and already measured quantity (e.g., length of tubed mercury) such that, if the measured quantity is arranged in order of its magnitude, the quantity to be measured is arranged in order of its magnitude as well. This is established through unbroken connection chains that establish proportional relations among the divisible properties of the qualitative properties connected (data generation traceability). These assumptions must be supported by experimental evidence through cross-validation (Chang, 2004; Ellis, 1966) and later also by theories about the underlying physical mechanisms and processes (e.g., temperature as kinetic energy of particles; Boon, 2015; Mari et al., 2015). Psychometricians tried to adapt these concepts to their "non-physical" study phenomena by *transforming the requirements of measurement into requirements of statistical assumptions*.

Psychometrics is not indirect (derived or associative) measurement

For empirical investigations, abstract conceptual entities such as constructs must be *operationally defined* in concrete entities that are directly accessible and thus (potentially) measurable (referents, indicators). This is often thought to be analogous to indirect physical measurement, thereby justifying the common assumption that quantifications for constructs could be derived from quantifications of some of their referents as is done in psychometric scaling. But as constructs are only construed and do not exist as real entities in themselves, construct indicators cannot be identified through experimentation as is possible for physical properties and their interconnections. Instead, scientists use specific theories, face validity or common-sense to assume that particular indicators are representative of a construct; this is called *fiat*[15] '*measurement'* (Cicourel, 1964). But the decreed links cannot be proven or be 'discovered' in nature because they are established through scientists' *interpretive decisions*. Operational definitions (e.g., scales) of the same construct therefore often vary.

Operationalization by decree is an unavoidable necessity for investigating constructs. But contrary to common beliefs, it is no step of measurement. To operationalize a construct, scientists specify the *qualitative* properties of study phenomena to which it is meant to refer. As abstractions, constructs typically refer to entities with *heterogeneous* qualities (e.g., different phenomena with different properties in which different aspects are emphasized). This makes it impossible to demarcate in constructs divisible entities of the *same* kind, thus *quantities* (Uher, 2020a). A hypothetical example may illustrate this. If one construed a 'weather' construct, operationalized with temperature, air pressure, humidity and wind strength, one could not identify divisible properties (quantities) in this 'weather' construct *in itself* but only in each of these specific indicators, featuring different qualities and different divisible properties. Such conglomerates of heterogeneous qualities (*blended concepts*; (Uher, 2018b) are, however, widely used in psychology. 'Extraversion', for example, refers to a conglomerate of various experiencings, behaviors, but also beliefs and attitudes, thus not only to different concrete

[15] From the Latin *fiat* for "let it be done".

phenomena each featuring different qualities but also to other conceptual phenomena. Such heterogeneous conceptual entities preclude the possibility to establish unbroken structural connections (causal links) to possible *quantitative* properties in the phenomena serving as their referents (construct indicators). 'Nice weather' cannot be causally linked, for example, to a temperature of 23 °C, an air pressure of 1023hPa, a relative humidity of 30% and a wind strength of 15 km/h. One may specify algorithms that define constellations of particular quantitative ranges for each of these qualitatively different weather indicators. But one cannot derive from them overall results to quantify 'weather' as is widely done, however, for psychological constructs like 'extraversion'. This highlights that, data generation traceability cannot be established from constructs as the actual objects of research in themselves (e.g., 'intelligence'; 'extraversion', 'weather'). This could be possible only for concrete indicators (referents) that are used as operationalizations (e.g., reaction times, task performances; physical properties used as indicators of 'weather')—an important point to consider when interpreting findings in construct research (for details, Uher, 2020a).

Fiat 'measurement' was also likened to metrologists' associative measurement (Torgerson, 1958). But the systematic structural relations of *quantitative* properties that are therefore required exist neither between psychical phenomena and the physical phenomena to which they are bound (e.g., brain physiology; Fahrenberg, 2013) nor between the entities of constructs and those of their (conceptual or concrete) construct referents. Unlike associative measurement, correspondence between the ordering of the assumed quantitative properties of constructs (e.g., 'intelligence') and the quantifications generated for their indicators (e.g., 'intelligence test' scores) cannot be shown (Trendler, 2013).

Psychometrics is not fundamental measurement

In *conjoint measurement theory* (Luce & Tukey, 1964), the additivity requirement of fundamental measurement, evidenced through empirical concatenation (Campbell, 1920), is transformed into a statistical requirement (Bond & Fox, 2003). This mathematical method allows to construct scales for research objects featuring multiple properties that are assumed to interact with one another and to jointly affect a property of interest (e.g., product properties influencing overall product evaluation). Scales are constructed such that the values of the target property are a function of the assumed values of the single component properties, thus allowing for the determination of optimal composition rules for scales. In additive conjoint measurement, a specific composition rule is assumed in order to develop scales with additive properties and units of equal distances. This enables the experimental testing of hypothesized additive structures with clear falsification criteria (Bond & Fox, 2003). This theory was praised for allowing to represent "meaningful and invariant amounts of *anything measured* in an additive, divisible, and portable numeric form" (italics added; Fisher, 2009, p. 1279). For this reason, it was considered a revolution in social-science measurement, fueling hopes it may finally enable the measurement of psychical 'attributes' (Michell, 1997, 1999).

This theory also underlies *Rasch models* (Rasch, 1980) that allow to transform nominal or ordinal manifest scores into interval scaled latent scores (Bond & Fox, 2003; Heene, 2013). The assumption that Rasch modelling—and latent variable modelling in general—could constitute measurement builds on the belief that the manifest data and the probabilities analyzed in them would constitute the empirical relational system under study (Borsboom & Scholten, 2008). But, as demonstrated above, this fundamental misconception derives from variable–referent and latent variable–construct conflations, numeral–number equation and the various misconceptions of semiotic systems in general. Specifically, probabilities are mathematical entities, not empirical entities (Kyngdon, 2008a). Psychometric models simply relate one numerical (symbolic) relational system to another instead of relating an empirical relational

system to a numerical[16] one as often (at least implicitly) assumed. It is for this reason that numerical values of the latent scale are not assigned by rules but instead *estimated through these models* (see Michell, 1997). Empirical relational systems, by contrast, involve "sets whose members are natural objects, events or relations" (Kyngdon, 2008a, p. 91). The term empirical, meaning experience-based (from Greek *empeiria* for experience), denotes the fact that the entities accessible to empirical study "will consist of only a restricted set of … objects or events, given the limitations of human cognitive and sensory-motor capacities" (Kyngdon, 2008a, p. 90-91; Uher, 2019).

Measurement—modelling confusion

It follows that psychometrics involves only data modelling but not data generation as required for measurement. This is also reflected in the facts that psychometric scales are the *outcome of* statistical data analysis and modelling, whereas *measurement scales* involve specific quantitative entities that are explicitly defined and agreed by convention *before* any measurement process is executed (BIPM, 2006). Statistics is not needed to develop measurement scales; physical measurement was successful long before statistics was developed (Abran et al., 2012). Indeed, Rasch models were developed from a mere mathematical framework (Rasch, 1980). "It is not necessarily wrong to develop mathematical models independently from empirical observations. But it is also not at all self-evident that empirical insights will result from such models" (Heene, 2013, p. 3). Measurement requires implementation of unbroken and traceable measurand–result connections (in RTM terms, systematic mapping relations between the empirical and the numerical relational structure), which involves testing the generated data for possible quantitative structures. But "in latent variable modelling in general, and Rasch modelling in particular, this is never done. Hence, Rasch enthusiasts are leaving something out and, according to Kyngdon [2008a], this something is not just important, but essential" (Borsboom & Scholten, 2008, p. 113).

Psychometricians produce continuous scales solely from statistical operations carried out on symbolic relational systems—yet without explaining, or at least aiming to explore, how the empirical relational systems that they assume to 'scale'—that is, psychical phenomena—can possibly vary that way. Neither Rasch models nor any other psychometric theories nor additive conjoint analysis provide a measurement framework in themselves (Trendler, 2009, 2013, 2019) and therefore enable neither fundamental nor any other kind of measurement. Psychometrics builds on a *fundamental confusion of measurement with statistical modelling*, thus of data generation with the analysis of data already generated (Figure 3a and b).

[16] Note that representation in a semiotic system (e.g., data) involves assignments not of numbers as sometimes assumed but of numerals the meaning of which must be first established empirically and through conventions.

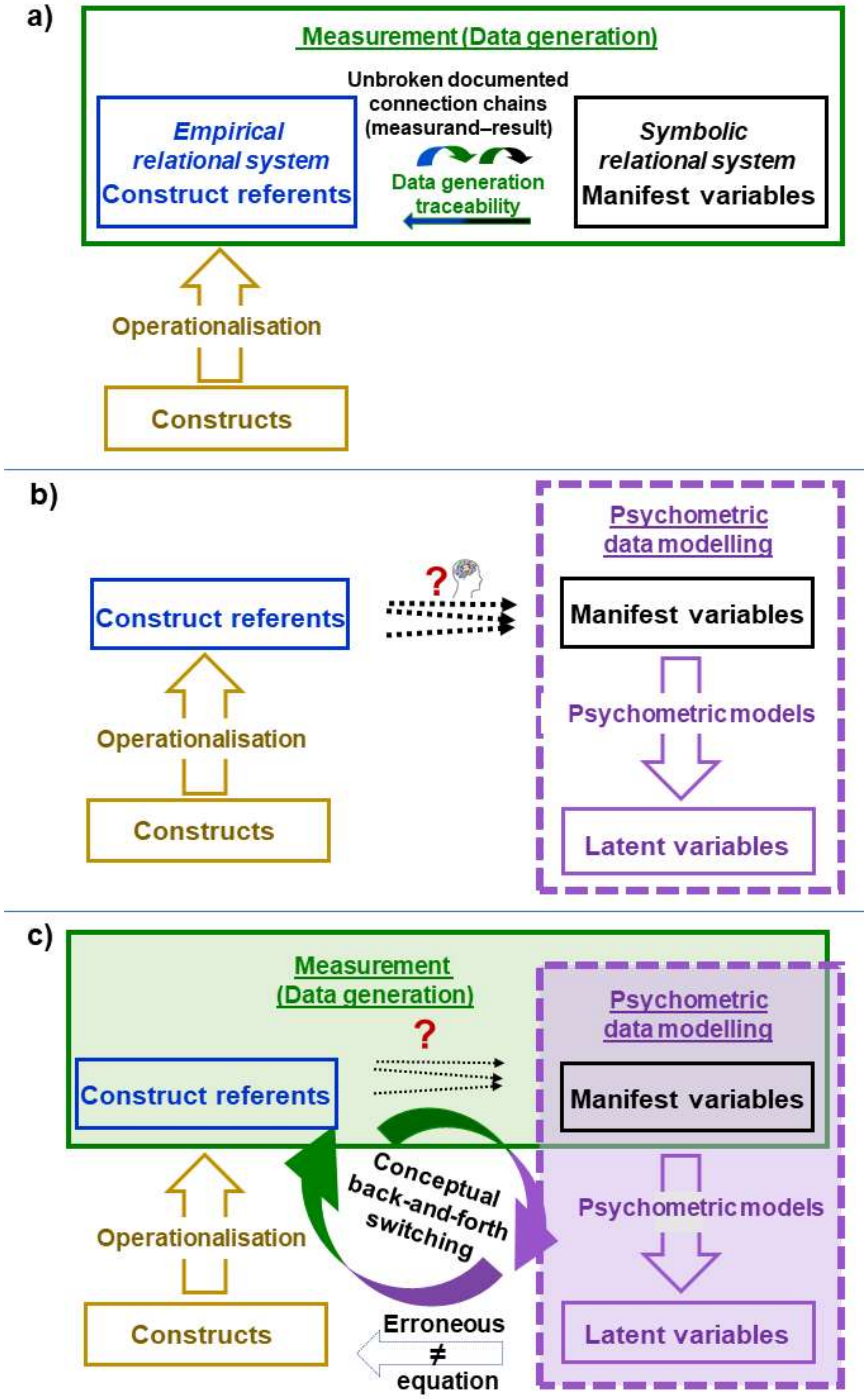

Figure 3. Two incompatible epistemological frameworks underlying psychometric theory and practice. *(a)* Realist measurement framework sporadically invoked but neither theoretically elaborated nor empirically implemented; *(b)* operationalist framework of data modeling well elaborated and empirically implemented; *(c)* conceptual flipping between both. The term variable means here semiotic encodings of the study phenomena but not the study phenomena in themselves.

This network of fallacies and psychometricians' focus on Steven's scale types misled many to believe that assumptions of quantitative properties in psychical phenomena could be empirically tested through statistical models that allow to create latent scales with such properties. Following this assumption, psychometricians develop statistical theories and models that allow to transform nominal or ordinal scaled scores on manifest (data) variables into overall scores on continuous latent (data) variables, which are assumed to be interval or even ratio scaled. Such properties, however, are properties of the latent (data) variables, which are mere statistical concepts. Latent (data) variables are neither the psychological constructs (abstract conceptual entities) they encode nor the psychical phenomena that these constructs may describe. "Latent variable models are not detectors of unobservable latent structures, properties/attributes, causal sources, or anything else" (Maraun & Halpin, 2008, p. 115). They are only mathematical-statistical models used to analyze data once these are generated. Thus, they are not models in the classical sense in that some real-world phenomena are modelled—i.e., represented in abstract ways to highlight structural patterns—because unbroken measurand–symbol connections are neither established nor even intended to be established (Kyngdon, 2008a; Maraun & Halpin, 2008). The quantitative properties identified in psychometric models can thus have only statistical and methodical origins, such as error structure (Michell, 2008), the probability concepts invoked (Kyngdon, 2008a), and the simplistic encoding format of rating scales aligned to statistical requirements rather than to divisible properties of the study phenomena (Uher, 2013, 2015d, 2018a).

Back-and-forth switching between two incompatible epistemological frameworks

This measurement—modelling confusion involves a repeated conceptual back-and-forth switching between two incompatible *epistemological frameworks—1)* an *operationalist framework of data modelling* that is well elaborated and empirically implemented and *2)* a *realist framework of measurement (data generation)* that is sporadically invoked in theoretical considerations but neither theoretically elaborated nor empirically implemented (Figure 3c). This switching is difficult to recognize because it is masked by the disparate notions of key terms like 'variables', 'traits' and 'attributes' and their unnoticed conflation. It occurs not only in the debates between proponents of different lines of thinking but also in the *same* authors' writings and often even in the *same* publication. This shows that this epistemological switching is inherent to the psychometric framework of thinking rather than reflecting only incompatibilities between different theoretical approaches.

At the bottom of the measurement—modelling confusion lies the recognition that, in lack of properties in psychical phenomena that are amenable to concatenation, "it is impossible to construct a representation on the basis of observed relations in the way this is done in axiomatic theory" (Borsboom & Mellenbergh, 2004, p. 112; Ferguson et al., 1940). This misled psychometricians to belief that representation would be unimportant for measurement. "[T]he fact that no representation theorem can be proven is not particularly disastrous. Of course, it would be nice to have one, but it is perfectly all right to speak of measurement if one does not" (Borsboom & Mellenbergh, 2004, p. 116). But any data generation, and thus any measurement framework, involves some kind of representation because data are signs that scientists construct to encode (signifier) information (meaning) from their study phenomena and properties (referents) for the purpose of analyzing the symbolic system thus-created *in lieu of* the empirical system under study.

Given the challenges for establishing representation theorems and the assumptions of their irrelevance, it is believed that "item response theorists must therefore take a philosophically important step: they have to assume, a priori, that a latent trait exists, and underlies our

observations" (p. 112). This statement, relying on the understanding of 'traits' as psychobiological phenomena existing in individuals, explicitly invokes the idea of a realist framework of measurement, thus contradicting the explicit belief representation would be unimportant. The realist assumption that phenomena and properties exist in the world and could potentially be made accessible to measurement has also driven early physicists and metrologists to develop pertinent theories and methodologies. But whereas many physical phenomena and properties were previously unknown and first had to be discovered (using suitable technologies) because they are *generally imperceptible* to humans, the existence of psychical phenomena is well-known because everyone can *perceive at least those that are consciously accessible in oneself*. Psychologists' challenges derive instead from the "non-physical" properties of psychical phenomena and their *fundamental imperceptibility in others* (Uher, 2016a). Hence, just assuming their existence does not answer the key question if and how measurement could be enabled (Michell, 2008) given their very special accessibility.

"Stevens (1946) brought relief by suggesting we restrict our research questions and the use of statistics *according to the structure of our measurement operations* instead of restricting ourselves to properties that demonstrably have additive structure. For this purpose he loosened the notion of measurement to 'the assignment of numerals according to rule', and defined his famous scales of measurement (nominal, ordinal, interval, and ratio) as dependent solely on the question which rule was followed in assigning the numerals" (italics added, Borsboom & Scholten, 2008, p. 111-112). Hence, psychologists focused on *methodical and statistical operations* that allow to produce additive structures in data. These operations involve, for example, recoding rating responses into numerals, their erroneous equation with numbers, operations of psychometric item analysis, iterative selection of only those item variables that allow to produce data with desired statistical properties while discarding those that do not (Uher, 2015d, 2018a) as well as operations of psychometric modelling of latent (data) variables that statistically explain variations in the manifest (data) variables (e.g., those used to define and estimate true scores; see Borsboom, 2005).

By focusing on these operations, psychometricians established a complex *operationist framework of data processing and modelling* that allows to produce statistically desired properties in empirical data. But none of these operations establish data generation traceability (unbroken measurand–result connections) and numerical traceability (linking the assigned numerical values to known standards) as needed for a framework of measurement. In psychometric modelling, systematic connections to the psychical phenomena studied in individuals are established *neither theoretically nor empirically*. Instead, quantitative structures are generated through specified operations of data processing (recoding) and statistical analysis (modelling). Interpretations of these quantitative structures are therefore bound to the particular operations used. "[O]perationism, as a philosophical theory of measurement, retains a discernable influence within psychology, for example, through S. S. Stevens's theory of measurement, and through its links to true score theory (Borsboom, 2005). Yet, operationism is deeply problematic and is widely considered to be philosophically unacceptable" (Haig & Borsboom, 2008, p. 3).

Psychometricians' operationist framework is detached from the question of how the data that are being processed and modelled were generated in the first place, which would require implementation of a realist framework of data generation and measurement. In rating methods, this difficult task is left to respondents who must construct for the symbolic relational system provided (e.g., item variables, scale categories) specific meanings and who must specify an empirical relational system to which they can relate these meanings as well as suitable assignments relations that enable structure-preserving mappings between both systems (e.g., homo-morphisms). But psychometricians commonly do not consider the raters' activities as the

actual data generation, likely because the conceptual switching between the disparate notions of 'variables' misleads scientists to conflate items describing the study phenomena with these phenomena in themselves (variable–referent conflation).

The measurement—data modelling confusion is also codified by the inconsistent use of the term 'causality', which contributes to the network of fallacies in psychometrics.

Disparate notions of 'causality' fortify the network of fallacies in psychometrics

The disparate notions of numerous key terms of psychological measurement (e.g., 'variables', 'attributes', 'responses') and their frequent conflation mislead psychologists to confuse the *unbroken causal connection chains* from the study phenomena's measurands to the results, needed to establish data generation traceability (and thus representation theorems), with the *psycho-bio-social causes* that underlie the emergence and functioning of the phenomena studied. This misconception is promoted by the frequent conflation of concrete phenomena in persons with the constructs (e.g., 'traits', 'intelligence') that scientists develop about them (construct–referent conflation). It is also fortified by psychologists' frequent reification of constructs as concrete entities residing inside persons (e.g., 'traits' as psychophysical mechanisms) that causally explain the concrete phenomena from which they were first derived (e.g., behaviors), which entails explanatory circularity (widespread e.g., in 'trait' psychology; Uher, 2013). Constructs (conceptual entities) are also often conflated with their encoding in latent (data) variables (latent variable–construct conflation). All this contributes to the interpretation of latent (data) variables as reflecting the phenomena (e.g., psychical processing) that may causally underly the observable 'responses' encoded in the data (e.g., test performances). But latent (data) variables underlying manifest (data) variables are mere statistical concepts used to formally or mathematically describe structures in numerical data, regardless of what kinds of referents these may encode.

This highlights that psychologists use the term 'causality' for at least three different relations, *1)* for measurand–result connections, *2)* for psycho-bio-social causes of target phenomena, and *3)* for statistical relations between (data) variables. The first notion refers to a realist measurement framework, the third to an operationist framework of statistical modelling. The second, however, requires neither theories of measurement nor of statistics but explanatory theories about the phenomena studied. The confusion of these disparate concepts of 'causality' entails a repeated switching between their very different notions, which are linked to different epistemological frameworks, resulting in leaps in argumentation and the confusion of measurement with data modelling.

Consequences

The network of fallacies on which the framework of psychometric theory and practice is based misleads psychometricians to belief that the quantitative properties produced in the latent (data) variables through various methodical and statistical operations would reflect properties of psychical phenomena as the actual phenomena of interest in themselves, thus providing evidence for their hypothesized quantitative structure. This erroneous conclusion is masked by a repeated (and commonly unnoticed) conceptual back-and forth switching between two incompatible epistemological frameworks, of which only the operationist framework is implemented, both theoretically and empirically. But without a (critical) realist framework of measurement, psychometrics cannot 'measure the mind' as the field aspires to do. Psychometric modelling, as its name indicates, involves only data modelling—thus data analysis—but not data generation as required for measurement. This highlights that the use of the term 'measurement'

in psychometrics is erroneous and unrelated to that established in metrology, the physical and other sciences. This constitutes a serious cross-scientific jingle-fallacy that, given the high public trust that our societies place on scientific measurement (Porter, 1995), has far reaching consequences for science and applied fields.

Inferences on psychical phenomena and thus persons inevitably biased

The radical alignment of psychometric modelling and instrument development to statistical assumptions and desired data structures rather than to the properties of the actual study phenomena entails that, despite all statistical sophistication, psychometric results can reveal only biased information about psychical phenomena, if any information at all (Toomela, 2008). As a consequence of this, it has "little real-world practical or even scientific consequence. For those who might doubt this, try and name just one result from the application of psychometrics in the last 50 years which has yielded a finding so important that it has changed the course of investigation and understanding of a phenomenon" (Barrett, 2008, p. 80). The focus on scaling techniques seems to have misguided psychometricians to overlook that creating scales is just a means to an end. "The Kelvin scale is a scientific achievement of first order, not merely because it provides a scale with mathematically powerful properties, but because it incorporates a profound understanding of how a certain class of phenomena works" (Duncan, 1984, p. 149). The aim of developing an understanding of the empirical study phenomena got out of sight in psychometrics.

These insights, reinforcing those already made by other scholars before (see above), demand much more careful interpretation of psychometric scores than is currently the case. They contribute theoretical justification for the increasing scrutiny that practitioners are placing on psychometric scores, such as on their validity as legal evidence in courts. It can no longer be justified that decisions on the application of the death penalty for offenders rests—even if just in part—on psychometrically determined IQ scores expressed to two-decimal place precision (Barrett, 2018), given that measurement is not even involved at all. It is just a matter of time that psychometric scores will be challenged in courts, like forensic psychologists' and psychiatrists' diagnostic practices before (Barrett, 2018; Faust, 2012).

Generating data about psychical phenomena: Key challenges

Psychometricians have not yet succeeded in establishing a measurement framework. Instead, in the widely-used rating methods, this crucial task is left to respondents' intuitive decisions, which are still hardly known, despite their intense use for almost a century now (Thurstone, 1928). This does not mean that rating data would be completely unrelated to the actual phenomena of interest; ultimately, ratings rest on common-sense knowledge and everyday words. But these are insufficient for developing scientific knowledge. Given the popularity of rating scales for creating quantitative data, it surprises how little has been invested in developing theories and empirical knowledge about how raters actually generate these data. Psychometricians neither provide nor build on pertinent theories (e.g., models of lexical stimulus perception and interpretation, theories about mental weighting of aspects considered for making decisions (Heene, 2013; Kyngdon, 2011; Uher, 2018a) as needed for establishing traceable data generation processes, and thus measurement. This also derives from the insufficiencies and vagueness of RTM and the failed implementation of at least a basic representation theorem in standardized rating methods. In lack of investigation of how raters actually interpret and use assessment scales, the serious violations of even the most basic preconditions of measurement in rating methods remained so far largely undetected.

Data generation methods capturing reaction times and 'responses' with specified meanings (e.g., correctness in 'intelligence' tests), by contrast, do enable the establishment of measurement processes. But this is possible only for these observable outcomes and not for the underlying psychical phenomena and processes as the actual phenomena of interest in themselves and one-to-one inferences to these latter are not possible. This is still not well considered in common interpretations of pertinent data (e.g., when 'intelligence' tests are interpreted as measuring 'intelligence' rather than intellectual performances). More efforts are needed to conceptually differentiate the phenomena involved (e.g., psychical versus behavioral phenomena). This involves methods that allow to implement both data generation traceability and numerical traceability, while carefully considering the peculiarities of psychical phenomena and thus inherent limitations. These two methodological principles, underlying structured frameworks of measurement established in metrology, are essential to ensuring robustness, replicability and usefulness of measurement data in all sciences. They are needed to enable public scrutiny and transparency, and to maintain a high degree of interpretability regarding study phenomena in the real-world and their properties.

The key challenge for psychological methods of data generation, however, remains the important distinction of the study phenomena in themselves from the means (e.g., concepts, models, methods, terms and data) used for their exploration. This article showed that mastering this distinction is far more intricate than it may seem at first sight. It requires the development of differentiated conceptual frameworks and an unambiguous terminology[17] in order to break up and overcome the network of fallacies and conflations that are codified in the measurement jargon currently established in psychology.

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). The author thanks Karen Slaney and Jim Cresswell for their editorial work as well as Manfred Schmitt, Andrew Maul and two anonymous reviewers for helpful comments on previous drafts.

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

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

Barrett, P. (2008). The consequence of sustaining a pathology: Scientific stagnation— a commentary on the target article "Is psychometrics a pathological science?" by Joel Michell. *Measurement: Interdisciplinary Research and Perspectives*, *6*(1–2), 78–83. https://doi.org/10.1080/15366360802035521

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

Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the Science of Self-Reports and Finger Movements: Whatever Happened to Actual Behavior? *Perspectives on Psychological Science*, *2*(4), 396–403. https://doi.org/10.1111/j.1745-6916.2007.00051.x

Berglund, B., Rossi, G. B., Townsend, J. T., & Pendrill, L. (2012). *Measurement with persons : theory,*

[17] Developing such frameworks and terminologies with a particular focus on research on individuals and involving various disciplines is a core aim of the TPS-Paradigm (Uher, 2015c, 2016a, 2019).

methods, and implementation areas. New York: Taylor Francis.
Bhaskar, R., & Danermark, B. (2006). Metatheory, Interdisciplinarity and Disability Research: A Critical Realist Perspective. *Scandinavian Journal of Disability Research*, *8*(4), 278–297. https://doi.org/10.1080/15017410600914329
BIPM. (2006). *BIPM: The international system of units (SI) (8th ed)*. Organisation Intergouvernementale de la Convention du Mètre. Retrieved from http://www.bipm.org/
Bohr, N. (1937). Causality and complementarity. *Philosophy of Science*, *4*(3), 289–298.
Bond, T. G., & Fox, C. M. (2003). Applying the Rasch Model: Fundamental measurement in the human sciences. *Journal of Educational Measurement*, *40*(2), 185–187. https://doi.org/10.1111/j.1745-3984.2003.tb01103.x
Boon, M. (2015). The scientific use of technological instruments. In S. O. Hansson (Ed.), *The role of technology in science: Philosophical perspectives* (Philosophy, pp. 55–79). Springer. https://doi.org/10.1007/978-94-017-9762-7_4
Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511490026
Borsboom, D. (2008). Latent Variable Theory. *Measurement: Interdisciplinary Research and Perspectives*, *6*(1–2), 25–53. https://doi.org/10.1080/15366360802035497
Borsboom, D., Cramer, A. O. J., Kievit, R. A., Scholten, A. Z., & Franić, S. (2009). The end of construct validity. In *The concept of validity: Revisions, new directions, and applications.* (pp. 135–170). IAP Information Age Publishing.
Borsboom, D., & Mellenbergh, G. J. (2004). Why psychometrics is not pathological. *Theory & Psychology*, *14*(1), 105–120. https://doi.org/10.1177/0959354304040200
Borsboom, D., & Mellenbergh, G. J. (2007). Test validity in cognitive assessment. In J. Leighton & M. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 85–116). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511611186.004
Borsboom, D., & Scholten, A. Z. (2008). The Rasch model and conjoint measurement theory from the perspective of psychometrics. *Theory & Psychology*, *18*(1), 111–117. https://doi.org/10.1177/0959354307086925
Campbell, N. R. (1920). *Physics: The Elements*. Cambridge, UK: Cambridge University Press.
Campbell, N. R. (2020). *Foundations of science*. Frankfurt am Main: Salzwasser Verlag.
Chakravartty, A. (2017). Scientific realism. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/archives/sum2017/entries/scientific-realism
Chang, H. (2004). *Inventing temperature: measurement and scientific progress*. Oxford University Press.
Cicourel, A. (1964). *Method and measurement in sociology*. New York: The Free Press of Glencoe.
Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957
Danziger, K. (1997). *Naming the mind: How psychology found its language*. London, UK: Sage.
Duncan, O. D. (1984). *Notes on social measurement: Historical and critical*. Nova York: Russell Sage Foundation.
Ellis, B. (1966). *Basic concepts of measurement*. Cambridge, UK: Cambridge University Press.
Epskamp, S., Rhemtulla, M., & Borsboom, D. (2017). Generalized network psychometrics: Combining network and latent variable models. *Psychometrika*, *82*(4), 904–927. https://doi.org/10.1007/s11336-017-9557-x
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]*. Lengerich, Germany: Pabst Science Publishers.
Faust, D. (2012). *Ziskin's coping with psychiatric and psychological testimony*. Oxford University Press. https://doi.org/10.1093/med:psych/9780195174113.001.0001
Ferguson, A., Myers, C. S., Bartlett, R. J., Banister, H., Bartlett, F. C., Brown, W., … Tucker, W. S. (1940). Quantitative estimates of sensory events: Final report of the committee appointed to consider and report upon the possibility of quantitative estimates of sensory events. *Advancement of Science*, *1*, 331–349.
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

Guyon, H., Falissard, B., & Kop, J.-L. (2017). Modeling psychological attributes in psychology – An epistemological discussion: Network analysis vs. latent variables. *Frontiers in Psychology*. Retrieved from https://www.frontiersin.org/article/10.3389/fpsyg.2017.00798

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

Hanfstingl, B. (2019). Should we say goodbye to latent constructs to overcome replication crisis or should we take into account epistemological considerations? *Frontiers in Psychology*, *10*, 1949. https://doi.org/10.3389/fpsyg.2019.01949

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] (3. Aufl.)*. Berlin: Walter de Gruyter.

Heene, M. (2013). Additive conjoint measurement and the resistance toward falsifiability in psychology. *Frontiers in Psychology*, *4*, 246. https://doi.org/10.3389/fpsyg.2013.00246

Heisenberg, W. (1927). Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik [On The Actual Content of Quantum Theoretical Kinematics and Mechanics]. *Zeitschrift Für Physik*, *43*(3–4), 172–198. https://doi.org/10.1007/BF01397280

JCGM200:2012. (2012). *International vocabulary of metrology – Basic and general concepts and associated terms (VIM 3rd edition)*. Working Group 2 (Eds.), Joint Committee for Guides in Metrology. Retrieved from https://www.bipm.org/utils/common/documents/jcgm/JCGM_200_2012.pdf

Khanam, S. A., Liu, F., & Chen, Y.-P. P. (2019). Comprehensive structured knowledge base system construction with natural language presentation. *Human-Centric Computing and Information Sciences*, *9*(1), 23. https://doi.org/10.1186/s13673-019-0184-7

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

Kuhn, T. S. (1962). *The structure of scientific revolutions*. Chicago: University of Chicago Press.

Kyngdon, A. (2008a). The Rasch model from the perspective of the representational theory of measurement. *Theory & Psychology*, *18*(1), 89–109. https://doi.org/10.1177/0959354307086924

Kyngdon, A. (2008b). Treating the pathology of psychometrics: An example from the comprehension of continuous prose text. *Measurement: Interdisciplinary Research and Perspectives*, *6*(1–2), 108–113. https://doi.org/10.1080/15366360802035570

Kyngdon, A. (2011). Plausible measurement analogies to some psychometric models of test performance. *British Journal of Mathematical and Statistical Psychology*, *64*(3), 478–497. https://doi.org/10.1348/2044-8317.002004

Lewin, K. (1936). *Principles of topological psychology*. New York, NY: McGraw-Hill.

Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology*, *1*(1), 1–27. https://doi.org/10.1016/0022-2496(64)90015-X

Lundmann, L., & Villadsen, J. W. (2016). Qualitative variations in personality inventories: subjective understandings of items in a personality inventory. *Qualitative Research in Psychology*, *13*(2), 166–187. https://doi.org/10.1080/14780887.2015.1134737

Maraun, M. D. (1998). Measurement as a normative practice. *Theory & Psychology*, *8*(4), 435–461. https://doi.org/10.1177/0959354398084001

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

Maraun, M. D., & Halpin, P. F. (2008). Manifest and latent variates. *Measurement: Interdisciplinary Research and Perspectives*, *6*(1–2), 113–117. https://doi.org/10.1080/15366360802035596

Maraun, M. D., Slaney, K. L., & Gabriel, S. M. (2009). The Augustinian methodological family of psychology. *New Ideas in Psychology*, *27*(2), 148–162. https://doi.org/10.1016/J.NEWIDEAPSYCH.2008.04.011

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

Mari, L., Carbone, P., & Petri, D. (2015). Fundamentals of hard and soft measurement. In A. Ferrero, D. Petri, P. Carbone, & M. Catelani (Eds.), *Modern measurements: Fundamentals and applications* (pp. 203–262). Hoboken, NJ: John Wiley & Sons. https://doi.org/10.1002/9781119021315.ch7

Maul, A. (2013). On the ontology of psychological attributes. *Theory & Psychology*, *23*(6), 752–769. https://doi.org/10.1177/0959354313506273

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

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

Michell, J. (2008). Is psychometrics pathological science? *Measurement: Interdisciplinary Research & Perspective*, *6*(1–2), 7–24. https://doi.org/10.1080/15366360802035489

Michell, J. (2012). Alfred Binet and the concept of heterogeneous orders. *Frontiers in Psychology*, *3*, 261. https://doi.org/10.3389/fpsyg.2012.00261

Narens, L. (2002). A meaningful justification for the representational theory of measurement. *Journal of Mathematical Psychology*, *46*, 746–768.

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, C. Hartshorne & P. Weiss (eds.), vols. 7-8, A. W. Burks (ed.)*. Cambridge, MA: Harvard University Press.

Pendrill, L. (2014). Man as a measurement instrument. *NCSLI Measure*, *9*(4), 24–35. https://doi.org/10.1080/19315775.2014.11721702

Pinheiro, M. A. (2020). A Wittgensteinian comment on “Psychology: A giant with feet of clay” A question from research on creativity. *Integrative Psychological and Behavioral Science*. https://doi.org/10.1007/s12124-020-09544-1

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

Rasch, G. (1980). *Probabilistic model for some intelligence and achievement tests*. Chicago, IL: University of Chicago Press.

Rorty, R. (1995). *Contingency, irony and solidarity*. Cambridge, UK: Cambridge University Press.

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

Salvatore, S. (2019). Beyond the meaning given. The meaning as explanandum. *Integrative Psychological and Behavioral Science*, *53*(4), 632–643. https://doi.org/10.1007/s12124-019-9472-z

Schacter, D. L., & Addis, D. R. (2007). Constructive memory: The ghosts of past and future. *Nature*, *445*(7123), 27–27. https://doi.org/10.1038/445027a

Slaney, K. L. (2017). *Validating psychological constructs: Historical, philosophical, and practical dimensions*. London, UK: Palgrave Macmillan. https://doi.org/10.1057/978-1-137-38523-9

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

Stent, G. S. (1969). The coming of the Golden Age: A view of the end of progress. Garden City: The Natural History Press. (The American Museum of Natural History).

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*, 667–680.

Thelen, E., & Smith, L. B. (1994). *A dynamic systems approach to the development of cognition and action*. Cambridge, MA: MIT Press.

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, *33*, 529–554.

Toomela, A. (2008). Variables in psychology: A critique of quantitative psychology. *Integrative Psychological & Behavioral Science*, *42*(3), 245–265. https://doi.org/10.1007/s12124-008-9059-6

Torgerson, W. S. (1958). *Theory and methods of scaling*. New York, NY: Wiley.

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 of ignoramus et ignorabimus? A rejoinder. *Theory & Psychology*, *23*(5), 591–615. https://doi.org/10.1177/0959354313490451

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

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). 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). New York: Springer International Publishing. https://doi.org/10.1007/978-3-319-10130-9_13

Uher, J. (2015b). 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 (Vol. 2), Series Neuropsychology: An interdisciplinary approach* (pp. 223–284). Berlin: Lit Verlag. https://doi.org/10.13140/RG.2.1.3848.8169

Uher, J. (2015c). 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. (2015d). 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. (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). New York: 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/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). London, UK: 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 & Quantity. International Journal of Methodology*, *53*(1), 221–246. https://doi.org/10.1007/s11135-018-0744-3

Uher, J. (2020a). 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. (2020b). Psychology’s status as a science: Peculiarities and intrinsic challenges. Moving beyond its current deadlock towards conceptual integration. *Integrative Psychological and Behavioral Science*. https://doi.org/10.1007/s12124-020-09545-0

Uher, J. (2021). 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., 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., & 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. (2012). *A guided science: History of psychology in the mirror of its making*. New Brunswick, NJ: Transaction Publishers.

Van Fraassen, B. C. (2012). Modeling and measurement: The criterion of empirical grounding. *Philosophy of Science*, *79*, 773–84.

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

von Glasersfeld, E. (1991). Knowing without metaphysics: Aspects of the radical constructivist position. In F. Steier (Ed.), *Research and reflexivity* (pp. 12–29). London: Sage.

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

Whitehead, A. N. (1929). *Process and reality*. New York, NY: Harper.

Wittgenstein, L. (2009). *Philosophical investigations*. Oxford, UK: Blackwell Publishing.

Woodward, J. F. (2011). Data and phenomena: a restatement and defense. *Synthese*, *182*(1), 165–179. https://doi.org/10.1007/s11229-009-9618-5

Wundt, W. (1896). *Grundriss der Psychologie [Outlines of Psychology]*. Stuttgart: Körner.

Wundt, W. (1907). *Logik der exakten Wissenschaften [Logic of the exact sciences], Band II (3. umgearb. Aufl.)*. Stuttgart: Enke.

***