Routledge
Taylor & Francis Group

Philosophical Psychology

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/cphp20

Causal complexity and psychological measurement

Markus Ilkka Eronen

To cite this article: Markus Ilkka Eronen (04 Jan 2024): Causal complexity and psychological measurement, Philosophical Psychology, DOI: 10.1080/09515089.2023.2300693

To link to this article: https://doi.org/10.1080/09515089.2023.2300693

Published online: 04 Jan 2024.

PHILOSOPHICAL PSYCHOLOGY
https://doi.org/10.1080/09515089.2023.2300693

Routledge
Taylor & Francis Group

OPEN ACCESS Check for updates

Causal complexity and psychological measurement

Markus Ilkka Eronen

Department of Theoretical Philosophy, University of Groningen, Groningen, The Netherlands

ABSTRACT
Psychological measurement has received strong criticism throughout the history of psychological science. Nevertheless, measurements of attributes such as emotions or intelligence continue to be widely used in research and society. I address this puzzle by presenting a new causal perspective to psychological measurement. I start with assumptions that both critics and proponents of psychological measurement are likely to accept: a minimal causal condition and the observation that most psychological concepts are ill-defined or ambiguous. Based on this, I argue that psychological measurement is fundamentally different from measurement in the physical sciences but can nevertheless be useful.

ARTICLE HISTORY
Received 5 September 2023
Accepted 26 December 2023

KEYWORDS
Measurement; causation; complexity; validity

1. Introduction

Psychological measurement has been a topic of much debate ever since the birth of psychological science. Measurements of attributes such as emotions, well-being, or intelligence are widely used for various purposes in society, but it remains a matter of discussion whether psychological measurement is analogous to measurement in the natural sciences, and to what extent it qualifies as measurement at all. Authors such as Michell (1997, 1999, 2008) and Trendler (2009, 2019) have forcefully argued that psychological researchers have not shown that psychological attributes are quantitative and that there is no reason to believe that psychologists are actually measuring anything. This criticism has led to much discussion among theoretical psychologists and philosophers (e.g., Borsboom & Mellenbergh, 2004; Bringmann & Eronen, 2016; Franz, 2022; Humphry, 2013; Sherry, 2011) but has had few if any practical implications. In practice, psychological measurement continues to play a fundamental role in both science and society. This results in a puzzling situation: Are the criticisms of

CONTACT Markus Ilkka Eronen markusilkka@gmail.com Department of Theoretical Philosophy, University of Groningen, Oude Boteringestraat 52, Groningen 9712 GL, The Netherlands

psychological measurement fundamentally flawed or are psychologists indeed not measuring anything, in spite of appearances?

In this paper, I present a new perspective to the debate, which explains both what the skeptical accounts of psychological measurement get right, and how psychological measurement practices can nevertheless be useful. My general approach is to focus on the *causal* underpinnings of psychological measurement, and put forward a minimal causal condition that measurement should satisfy. This condition is so minimal and weak that both critics and proponents of psychological measurement will find it hard to deny. I then show that, in light of this condition, psychological measurement practices are not analogous to measurement in physics but can nevertheless be useful.

My approach assumes the interventionist account of causation: Due to its minimalistic nature and continuity with scientific practice and causal modeling, it provides an appropriate notion of causation for the purposes of this paper. The interventionist account has been developed by Woodward (2003) and reflects insights from the causal modeling literature (Pearl, 2009; Spirtes et al., 2000). Its core idea is that causes make a difference to their effects and that causal relationships are potentially or ideally exploitable for manipulation and control. More precisely, X is a cause of Y if and only if the value (or the probability distribution) of Y would change under some intervention on X, and the intervention has to change X in such a way that the change in Y is only due to the change in X and not any other factor (see Woodward, 2003, 2015 for more details). Importantly, interventionism does not require any metaphysical assumptions about the nature of causation or the causes themselves; all that is needed is some coherent conception of how the putative causes can be changed or manipulated.

The structure of this paper is as follows. First, I will introduce the minimal causal condition for measurement. Next, I discuss causal complexity and conceptual ambiguity in psychology, and how they (in conjunction with the minimal causal condition) lead to fundamental obstacles to psychological measurement. I then discuss possible objections, and conclude by considering the implications for psychological measurement and ways forward.

2. A minimal causal condition for measurement

Throughout the 20th century, causality did not play a significant role in discussions of psychological measurement. The reasons for this are complex, but one important factor is general hesitance and avoidance of causal language, largely due to the influence of logical positivism (Borsboom, 2005; Grosz et al., 2020). In the philosophy of science, discussions of measurement have revolved around the representational theory of measurement (RTM; Suppes et al., 1989). As its name suggests, RTM is focused on representational relationships between empirical and numerical systems, and does not refer to causal notions.

However, in recent times causality has started to find its way to discussions of psychological measurement (Bollen, 2002; Bollen & Pearl, 2013; Borsboom et al., 2004; van Bork et al., 2022). A key insight underlying these approaches is that in valid measurement the attribute or property that is being measured should *cause* the measurement outcome. To take an example from physics, in valid temperature measurements, variations in temperature should cause variations in thermometer readings. Similarly, in valid measurements of a psychological attribute such as sad mood, variations in sad mood should cause variations in responses to items measuring sad mood.[1]

More concretely, it has been argued that psychometric measurement models, more specifically latent variable models, should be seen as causal models (Bollen & Pearl, 2013; Markus & Borsboom, 2013; van Bork, Wijsen, et al., 2017). Latent variable models are at the core of state-of-the-art accounts of testing and measurement in psychology, such as item response theory (Markus & Borsboom, 2013). A standard latent variable model consists of one unobserved (latent) variable, such as reading proficiency, and several observed variables or indicators, such as variables representing responses to items in a questionnaire or test (e.g., a reading proficiency test). The values of the observed variables are seen as measurements of the latent variable, possibly with some error.

Measurement models of this kind are most plausibly interpreted as causal models that reflect causal assumptions (Bollen & Pearl, 2013; van Bork, Wijsen, et al., 2017). For example, the latent variable models standardly assume “local independence”, meaning that the observed variables are statistically independent of each other, when controlling for the latent variable. This is exactly the case when the latent variable is the common cause of the observed variables, and the observed variables do not causally interact with each other. Assumptions like this are hard to justify if the model is treated as a non-causal model, for example, as a summary statistic (Bollen & Pearl, 2013; van Bork, Wijsen, et al., 2017). Therefore, insofar as psychological measurement models are latent variable models, they seem to be (tacitly) assuming that the attribute that is measured is causing the measurement outcomes.

Borsboom and colleagues have further developed this insight into an even stronger and more explicitly causal account of psychological measurement (Borsboom, 2005; Borsboom et al., 2004). They take as a starting point the core idea of validity: a test or measurement is valid if and only if it measures what it is intended to measure. Based on this, they elevate causality to a necessary and sufficient condition for valid measurement: “a test is valid for measuring an attribute if and only if ... variations in the attribute *causally produce* variations in the outcomes of the measurement procedure” (Borsboom et al., 2004, p. 1061, emphasis added).

For the purposes of this paper, it is not necessary to defend a full-blown causal account of measurement. Instead, what is enough is the following minimal causal condition, which can be distilled from the accounts discussed above:

Minimal causal condition for measurement: *O* is a valid measure of *X only if* there is a causal relationship from *X* to *O*.

This minimal causal condition posits that a causal relationship between the target of measurement and the measurement outcome is a necessary (but not sufficient) condition for valid measurement. *X* can refer to an attribute, property, or a state; this condition is intended to be as minimal as possible, and compatible with various views of the nature of the targets of measurement. Similarly, *O* can refer to any outcome of a measurement procedure, such as instrument readings or responses to items in a questionnaire. Note also that the minimal condition does not require that we actually *know* what the causal relationship is, just that there *is* a causal relationship from *X* to *O*. The problem (that I will discuss in the following section) is that in psychology we usually do not have justification to assume the latter (i.e., that there is a causal relationship from *X* to *O*).

As this condition is only a necessary one, it is compatible with further requirements for valid measurements. For example, one might also require evidence that the attribute of interest is in fact a quantitative attribute (Michell, 1999, 2008). One might also add requirements from model-based approaches to measurement: For example, that one needs to provide a model of the actual measurement process and interactions of the instrument and the attribute of interest (Mari, 2005; Tal, 2017). These further conditions are compatible with but not directly releivant for my argument, as I will focus on the consequences that the minimal condition by itself has for psychological measurement.

3. Conceptual ambiguity and causal complexity

In spite of being a relatively weak assumption, the minimal causal condition results in severe problems for psychological measurement. I will focus on two features of psychology that, in conjunction with the minimal causal condition, lead to trouble: the conceptual unclarity that abounds in psychology, and the massive causal complexity of the mind-brain.

First, it is widely accepted that most psychological concepts are not well-defined (Flake, 2021; Franz, 2022; Podsakoff et al., 2016; Scheel et al., 2021). A salient example of this is psychological symptoms

(Bringmann et al., 2022; Wilshire et al., 2021). Symptoms are crucial for diagnosing and treating individuals with mental disorders, and also play a central role in explanatory models of mental disorders (such as network models, Borsboom, 2017). However, what symptoms are and how they should be conceptualized is far from straightforward. For example, as Wilshire and colleagues point out, patient narratives suggest that the symptom 'impaired concentration' may variously refer to experiencing a mental blank, being interrupted by intrusive thoughts, or drifting off topic. These seem like distinct phenomena: experiencing mental blanking is very different from suffering from intrusive thoughts. In other words, at closer inspection, the concept 'impaired concentration' seems to refer to several phenomena that are in fact distinct. Importantly, this kind of conceptual ambiguity is not unique to impaired concentration, but also holds for other symptoms such as anhedonia and fatigue (Wilshire et al., 2021).

From a causal perspective, this conceptual unclarity is detrimental to psychological measurement, because it means that we do not know the causal structure of concepts we are intending to measure. To see this, let us focus on impaired concentration as an example, and consider a situation where impaired concentration is measured with multiple items. This then corresponds to a standard psychometric measurement model, where impaired concentration is represented as a latent variable, and the observed variables represent responses to different items assessing impaired concentration. The problem is that if impaired concentration actually consists of several distinct phenomena, this representation is not correct (see also VanderWeele, 2022). Impaired concentration is not a single phenomenon, but actually refers to three distinct phenomena that are causally different (i.e., their causal relationships to other phenomena are different from each other) and therefore should be represented with distinct variables. Moreover, each of these three variables may be causally related to the observed variables in different ways: for example, perhaps blanking causes positive responses to variable A but not B, and drifting off topic causes positive responses to variable B but not A, and so on. Thus, this is not a case of multiple realizability, where the same property (impaired concentration) can be realized in different ways (e.g., as blanking and/or as intrusive thoughts). Rather, the property as it was originally conceived "fractionates" (Wilshire et al., 2021) into different properties that are causally distinct and only superficially similar.

The upshot is that when a standard measurement model represents impaired concentration as a single variable, whereas it in fact comprises three distinct phenomena, it is a causally incorrect model. This is where the minimal causal condition becomes relevant: In this case, the measurement outcomes (observed variables) are *not* caused by the variable we are intending to measure (impaired concentration), but by one of the three other phenomena. Therefore, the minimal causal condition is not satisfied, and we do not have measurements of impaired concentration.

If impaired concentration would be just an isolated example, this problem would not be very devastating for psychology. However, similar issues of conceptual unclarity have been identified for a broad range of other central psychological concepts, such as emotions (Weidman et al., 2017), well-being (Alexandrova, 2012, 2017), or attention (Hommel et al., 2019). In fact, it is hard to find a central psychological concept that has not received criticism for being ill-defined, ambiguous, or vague, and there are many overview articles laying bare the conceptual unclarity that abounds in psychology (e.g., Bringmann et al., 2022; Podsakoff et al., 2016; Scheel et al., 2021).

Thus, in order to make progress in psychological measurement, it seems that what is needed is a clearer conceptualization of psychological phenomena that actually reflects the underlying causal structure. However, this is hampered by the second problem I mentioned at the outset, namely causal complexity. The human brain is often called the most complex system in the known universe, consisting of around 87 billion neurons, and a single neuron may be connected to up to 100 000 other neurons (Yuste, 2015). The challenge is then to distill relatively stable and causally well-behaved variables, at various levels, from this nearly unfathomable complexity (see also Eronen, 2021; Trendler, 2009). There are nowadays more and more initiatives aiming to address this challenge (Francken et al., 2022; Poldrack & Yarkoni, 2016), but traditionally efforts in psychology and neuroscience have not been concentrated on this kind of conceptual and ontological work. Instead, the focus has been on applying statistical methods to empirical data, without much attention to how phenomena are conceptualized (Bringmann et al., 2022; Eronen & Romeijn, 2020).

The implication of this is that models based on *current* psychological concepts are likely to oversimplify and misrepresent the causal structure. For this reason, there is little reason to believe that current psychological concepts are the causes of measurement outcomes: just like in the case of impaired concentration, it is likely that the causal structure is far more complex. Therefore, there is no justification for assuming that measurements of psychological variables satisfy the minimal causal condition.

4. Possible objections

In response to the above, one could argue that the issue can be solved empirically: perhaps we could use statistical techniques to empirically study whether an attribute, such as impaired concentration, in fact, consists of one or several phenomena. Most saliently, factor analysis is a statistical method that has been very widely used in psychology to study how many latent factors best explain the observed data (Johnson, 2016). In the early days of scientific psychology, factor analysis was used to indicate that the positive correlations among scores of different types of intelligence tests, known as the "positive manifold", are best explained by one underlying factor, usually called the *g* factor (Van der Maas et al., 2006). Similarly, we could, for instance, check whether measurements of impaired concentration are best explained by one, two, three, or more latent variables.

However, it has been widely established that factor analysis is a poor guide to causal structure (Johnson, 2016; Romeijn & Williamson, 2018; van Bork, Epskamp, et al., 2017; VanderWeele, 2022). Even in cases where factor analysis strongly supports a certain solution, different causal structures can explain the data equally well, both in principle and in practice (van Bork, Epskamp, et al., 2017). Intelligence is a case in point: In recent year, van der Maas and colleagues have put forward an alternative theory of intelligence, where the positive manifold is explained by a complex dynamic system, with several cognitive components that causally interact through development (Van der Maas et al., 2006; Maas et al., 2017; see also VanderWeele, 2022). Although the debate is far from settled, this alternative model seems to explain the relevant phenomena at least as well as the traditional one-factor model, while doing more justice to the actual complexity of the mind-brain (see van der Maas et al., 2017 for discussion). However, for the present context, the most important point is that factor analysis is not sufficient to determine how many variables there are and what is the cause of the measurements.

The same holds for empirical or data-driven methods to infer causal structure, such as causal discovery algorithms (Spirtes et al., 2000). In order to give informative results, they either require substantial initial knowledge of the causal structure of the system of interest, which we do not currently have in psychology due to conceptual unclarity, or strong causal assumptions (such as causal sufficiency), which are not warranted in psychology due to the massive causal complexity (see also Eronen, 2020). To summarize, there is no magic bullet that would empirically solve the problems of conceptual unclarity and causal complexity.

A different pathway to avoid the negative conclusion regarding psychological measurement would be to simply reject the minimal causal condition. In order for the condition to fail, there should be cases of valid measurement where there is *no* causal relationship between the variable that is being measured and the measurement outcome. In other words (and somewhat simplified), there should be cases where *O* is a valid measure of *X*, but intervening on *X* is *not* a possible way to change *O*. However, it is hard to see how such a situation could still qualify as valid measurement of *X*. Perhaps one might argue that a strong correlation between *X* and *O*, in combination with further requirements (e.g., from the representational theory of measurement), is sufficient to conclude that *O* is a measure of *X*, even in the absence of a causal relationship. However, if *X* is not a cause of *O*, a reliable strong correlation between *O* and *X* is only possible if (1) *O* is cause of *X*, (2) *O* and *X* are both caused by a third variable (common cause), or (3) *O* and *X* are not distinct but conceptually connected (e.g., due to semantic overlap).

Options (1) and (3) can be easily dismissed: Measurements should not cause the thing that is being measured (1), and measurements should be conceptually distinct from the target of measurement (3). This leaves only (2), but also in this case, it is hard to see how such common causes structures could support measurement. Consider a situation where we are trying to measure a psychological variable *X* with variable *O*, and *X* and *O* are strongly correlated due to a common cause *C*, but there is no causal relationship from *X* to *O*. The problem is that if *X* varies due to some other factor than *C*, values of *O* will no longer track values of *X* at all, so "measurement" of this kind would be highly unstable and context-dependent. At best, it would only count as measurement in a much weaker sense than measurement in the physical sciences, a topic to which I return to in the next section. Moreover, in order to know whether *C* is present and acting as a common cause of *O* and *X*, we would again require some knowledge of the (psychological) causal structure (e.g., how *C* causes the psychological variable *X*), which would lead to the problems discussed in the previous section.

5. Implications for psychological measurement

The arguments in this paper are based on straightforward and widely accepted assumptions: The causal complexity of psychology and the observation that most psychological concepts are ill-defined or ambiguous. Together with the minimal causal condition, which I have defended in the previous section, they lead to the conclusion that psychological measurement, at least in most cases, does not fulfill the requirements for genuine measurement.

A similar conclusion, but based on different arguments, has been reached by several other authors in recent years, such as Franz (2022), Michell (1999, 2008), Trendler (2009, 2019), and Uher (2021). For example, Trendler argues that psychological phenomena "*are neither manipulable nor are they controllable to the extent necessary for an empirically meaningful application of measurement theory. Hence they are not measurable*" (Trendler, 2009, p. 592). Michell argues that psychologists have failed to show that the attributes measured are actually quantitative (Michell, 1999), and Uher describes a “complex network of fallacies” underlying psychological measurement practices (Uher, 2021). The arguments outlined in this paper could therefore simply be taken as further reasons supporting the negative conclusion regarding psychological measurement.

However, a problem with this conclusion is that the results of psychological measurements are often quite useful. As one example, a recent paper by an economist and a statistician provided an interesting outsider view on this issue: Sidestepping the earlier discussions on psychological measurement, these authors looked at the predictive value of “numerical measures of subjective feelings”, such as happiness scores based on self-reports. The authors concluded that such psychological measures are often better predictors of decisions and actions (e.g., changing job or moving to a new house) than economic variables such as socioeconomic status (Kaiser & Oswald, 2022). As a more classic example, intelligence test scores have been shown to have predictive value for various outcomes, such as job performance or academic achievement (see, e.g., Rohde & Thompson, 2007). These observations suggest that the results of psychological measurements can be informative and useful, at least for specific purposes.

The approach that I propose to account for this is to distinguish between (a) physics-style measurement, which for convenience we can call *hard measurement*, and (b) data generation by human participants, which we can call *soft measurement*. The former refers to measurement as it is understood in the physical sciences, metrology (the science of measurement), or the philosophy of measurement. As I have argued in this paper, the minimal causal condition is a necessary condition for measurement in this sense. The latter refers to practices such as participants filling in questionnaires on their mood states, or responding to items in an intelligence test. In these situations, human individuals are putting numbers on things, thereby generating psychological data. However, it is crucial that this is very different from hard measurement: we have no knowledge of the actual causal structure and what is causing the measurement outcomes and no justification for assuming a causal link between the attribute of interest and the measurement outcomes.

Making the distinction between hard and soft measurement has two key advantages. First, it can help to move the debate on psychological measurement forward by providing more nuance and common ground. The conclusion that psychological measurement is deeply problematic is hard to deny, but to take this further and argue that psychological measurement is not measurement at all (in line with Michell, Trendler, and others) is also problematic, considering that there is a whole branch of science devoted to psychological measurement (psychometrics), and those psychological measurements are often useful for specific purposes. Thus, a more nuanced conclusion is that psychological measurement is not *hard* measurement, but nevertheless measurement in a weaker sense. This is something that both proponents and critics of psychological measurement could accept, which may lead to much-needed common ground in the debate on the nature of psychological measurement.

Second, this distinction highlights the need to better understand soft measurement on its *own* terms, as something distinct from hard measurement. As I have argued above, psychometric measurement models, most importantly latent variable models, are most plausibly seen as causal models. Therefore, psychometrics tends to implicitly assume a hard causal picture of measurement (and authors such as Borsboom explicitly endorse it). In contrast, I have argued that psychological measurement should be conceived as soft measurement. This allows researchers to focus on the specific features of soft measurement, without trying to force psychological measurement into the mold of hard measurement.

To make this point clearer, it is helpful to consider hard and soft measurement in relation to validity, a central concept in debates on psychology measurement. The notion of validity is notoriously ambiguous and there is no agreement on how it should be understood, but among the many attempts to define it, we can identify two general strands that map well onto the distinction between hard and soft measurement. First, as discussed in section 2, Borsboom and colleagues argue that validity should be understood causally: "a test is valid for measuring an attribute if and only if … variations in the attribute causally produce variations in the outcomes of the measurement procedure" (Borsboom et al., 2004, p. 1061)". As I have argued, it is appropriate to require this kind of validity when it comes to hard measurement.

However, there is a different tradition of approaching validity, where no causality is required, but validity is rather seen as a matter of providing arguments and evidence for the intended use of measurements (e.g., Cook & Beckman, 2006; Kane, 1992, 2013). In this approach, validity is a matter of degree, and can only be assessed relative to a specific intended interpretation or use. For example, it is plausible that intelligence test scores have a degree of validity for predicting academic achievement. The evidence for this intended use mainly consists in correlations between intelligence test scores and measures of academic achievement (Rohde & Thompson, 2007). However, a degree of validity for this specific use does not mean that intelligence test scores are valid in a more general (causal) sense, or that they reflect an underlying latent variable that is a cause of academic achievement. In fact, in light of what I have argued in this paper, it is very likely that these additional claims are false: the causal structure leading to the measurement outcomes is unknown, and evidence points toward a complex system of interacting variables instead of the traditional latent variable picture (see section 4).[2] Thus, the argument-based approach to validity allows us to see how psychological measurements can have a degree of validity for certain intended uses or interpretations, such as prediction, even though they are not valid in the causal sense that corresponds to hard measurement.

It may seem odd that measurements (e.g., of intelligence) can be predictively useful even when the target of measurement is not actually the cause of the measurement outcomes and is also not causing the variable that is being predicted. However, prediction is possible purely based on (noncausal) correlations. For example, barometer readings can be used to predict storms, even though they are just correlated with them and not causing them. Similarly, psychological measurements can be predictive based on correlations (e.g., between intelligence test scores and academic achievement), also in the absence of a causal link. In addition, even concepts that are ambiguous or not well-defined can have a degree of usefulness for prediction, as long as they are correlated with some variables that are in fact causes of the variable that is being predicted. This is presumably the case with intelligence: Although 'intelligence' is an ambiguous and poorly defined concept, it is probably correlated with various (cognitive, affective, social, etc.) variables that are, through extremely complex pathways, causes of academic achievement.

In this way, conceiving psychological measurement as soft measurement allows us to see how psychological measurements can be useful without being causal. Moreover, it results in a shift in emphasis from measurement models and causal processes to the conceptual and qualitative aspects of measurement.[3] When we approach intelligence measurement from the perspective of soft measurement, it becomes clear that we need to carefully consider what the intended uses of the measurements are, what the evidence for them consists in, and what conceptualization of 'intelligence' is assumed (e.g., is intelligence seen as a purely statistical construct, as a complex dynamic system of interacting components, or as something else?). In the case of self-report-based measurements of variables such as 'happiness' or 'impaired concentration', questions that become central include: What kind of conceptualization of happiness or concentration in implicit in the questionnaire? Is this the right conceptualization for the intended use or interpretation? How could we make the concept less ambiguous and better suited for the intended use or interpretation? Answering these questions clearly requires conceptual work, such as disambiguation, explication, and deliberation between different stakeholders, and is not something that can be done by psychometrics alone (see Alexandrova, 2017). In addition, if we want to understand how participants come up with their answers to items in a questionnaire or test, and whether this matches with the assumptions of researchers or other stakeholders, we also need qualitative research, such as interviewing participants (e.g., cognitive interviewing) or studying their open box answers (see, e.g., Beatty & Willis, 2007; Castillo-Díaz & Padilla, 2013).

All of this is in stark contrast to mainstream psychology, where measurement in general does not receive much attention, and when it does, the focus is on its statistical and quantitative aspects (Alexandrova, 2017; Flake et al., 2017; Fried & Flake, 2018). Approaching soft measurement from various different angles, including not only statistical tools (e.g., assessing reliability and calculating different types of correlations) but also the above-mentioned conceptual and qualitative approaches, will help to make psychological measurements more robust and more fruitful for the intended uses.

As a final point, it is important to emphasize that these problems are not unique to psychology. They stem from causal complexity and conceptual unclarity, which will also be familiar to researchers from other fields. However, contemporary psychology does provide a very clear and extreme case, both regarding conceptual unclarity (Bringmann et al., 2022; Wilshire et al., 2021) and causal complexity (Eronen, 2021; Pessoa, 2023). One can see the severity of these issues as a continuum, where at the one end we find fields such as clinical psychology, and at the other end cases in physics where causal connection from the attribute of interest to the measurement outcome is well understood (e.g., length, mass, and luminous intensity). Thus, the issues discussed in this paper have broader relevance, but they are particularly crucial in the context of psychology, especially considering the societal importance of psychological measurements.

Notes

1. It is a matter of debate what validity exactly consists in, but the basic idea is that in valid measurement the test or instrument is measuring what it is intended to measure. Although I will not always make it explicit, I will focus in this paper on validity instead of other desiderata for measurement (e.g., reliability). I will also explicitly discuss different notions of validity and their relationship to measurement in section 5.
2. In this light, we should be very skeptical regarding any claims that attribute intelligence a causal role based on intelligence measurements. This is particularly important considering the morally problematic and controversial history of intelligence measurement (see, e.g., Serpico, 2021). In addition, even if intelligence scores are useful for some specific purposes such as prediction, the harms of using them may outweigh the benefits; see, e.g., Kourany (2020) and Uchiyama et al. (2021) for critical discussion.
3. Importantly, this is not an either-or question: quantitative methods (e.g., calculating reliability or correlations) can be useful for soft measurement, and conceptual clarification and iterative refinement of concepts are also crucial for hard measurement (Bringmann & Eronen, 2016; Chang, 2004; Eronen & Romeijn, 2020).

Acknowledgement

I would like to thank Alessandra Basso, Laura Bringmann, Cristian Larroulet Philippi, Freek Oude Maatman and Jani Raerinne for their helpful comments on earlier drafts of this paper, as well as two anonymous reviewers of this journal for their constructive and insightful reviews. Previous versions of this paper were presented at the PSA 2022 in Pittsburgh, the TINT Perspectives on Science seminar in Helsinki, and the PCCP colloquium in Groningen; I am grateful to the audiences for their feedback. Finally, special thanks to Denny Borsboom, with whom I have discussed these ideas on many occasions, and whose work provided the starting point and foundation for this paper.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

Alexandrova, A. (2012). Well-being as an object of science. *Philosophy of Science*, *79*(5), 678–689. https://doi.org/10.1086/667870

Alexandrova, A. (2017). *A philosophy for the science of well-being*. Oxford University Press.

Beatty, P. C., & Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly*, *71*(2), 287–311. https://doi.org/10.1093/poq/nfm006

Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology*, *53*(1), 605–634. https://doi.org/10.1146/annurev.psych.53.100901.135239

Bollen, K. A., & Pearl, J. (2013). Eight myths about causality and structural equation models. In S. L. Morgan (Ed.), *Handbook of causal analysis for social research* (pp. 301–328). Springer Netherlands. https://doi.org/10.1007/978-94-007-6094-3_15

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge University Press. https://doi.org/10.1017/CBO9780511490026

Borsboom, D. (2017). A network theory of mental disorders. *World Psychiatry*, *16*(1), 5–13. https://doi.org/10.1002/wps.20375

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

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

Bringmann, L. F., Elmer, T., & Eronen, M. I. (2022). Back to basics: The importance of conceptual clarification in psychological science. *Current Directions in Psychological Science*, *31*(4), 340–346. https://doi.org/10.1177/09637214221096485

Bringmann, L. F., & Eronen, M. I. (2016,). Heating up the measurement debate: What psychologists can learn from the history of physics. *Theory and Psychology*, *26*(1), 27–43. https://doi.org/10.1177/0959354315617253

Castillo-Díaz, M., & Padilla, J.-L. (2013). How cognitive interviewing can provide validity evidence of the response processes to scale items. *Social Indicators Research*, *114*(3), 963–975. https://doi.org/10.1007/s11205-012-0184-8

Chang, H. (2004). *Inventing temperature: Measurement and scientific progress*. Oxford University Press.

Cook, D. A., & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: Theory and application. *The American Journal of Medicine*, *119*(2), 166.e7–16. https://doi.org/10.1016/j.amjmed.2005.10.036

Eronen, M. I. (2020). Causal discovery and the problem of psychological interventions. *New Ideas in Psychology*, *59*, 100785. https://doi.org/10.1016/j.newideapsych.2020.100785

Eronen, M. I. (2021). The levels problem in psychopathology. *Psychological Medicine*, *51*(6), 927–933. https://doi.org/10.1017/S0033291719002514

Eronen, M. I., & Romeijn, J.-W. (2020). Philosophy of science and the formalization of psychological theory. *Theory & Psychology*, *30*(6), 786–799. https://doi.org/10.1177/0959354320969876

Flake, J. K. (2021). Strengthening the foundation of educational psychology by integrating construct validation into open science reform. *Educational Psychologist*, *56*(2), 132–141. https://doi.org/10.1080/00461520.2021.1898962

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, *8*(4), 370–378. https://doi.org/10.1177/1948550617693063

Francken, J. C., Slors, M., & Craver, C. F. (2022). Cognitive ontology and the search for neural mechanisms: Three foundational problems. *Synthese*, *200*(5), 378. https://doi.org/10.1007/s11229-022-03701-2

Franz, D. J. (2022). 'Are psychological attributes quantitative?' Is not an empirical question: Conceptual confusions in the measurement debate. *Theory & Psychology*, *32*(1), 131–150. https://doi.org/10.1177/09593543211045340

Fried, E. I., & Flake, J. K. (2018). Measurement Matters. *APS Observer*, *31*(February). https://www.psychologicalscience.org/observer/measurement-matters

Grosz, M. P., Rohrer, J. M., & Thoemmes, F. (2020). The taboo against explicit causal inference in nonexperimental psychology. *Perspectives on Psychological Science*, *15*(5), 1243–1255. https://doi.org/10.1177/1745691620921521

Hommel, B., Chapman, C. S., Cisek, P., Neyedli, H. F., Song, J.-H., & Welsh, T. N. (2019). No one knows what attention is. *Attention, Perception, & Psychophysics*, *81*(7), 2288–2303. https://doi.org/10.3758/s13414-019-01846-w

Humphry, S. M. (2013). A middle path between abandoning measurement and measurement theory. *Theory & Psychology*, *23*(6), 770–785. https://doi.org/10.1177/0959354313499638

Johnson, K. (2016). Realism and uncertainty of unobservable common causes in factor analysis. *Noûs*, *50*(2), 329–355. https://doi.org/10.1111/nous.12075

Kaiser, C., & Oswald, A. J. (2022). The scientific value of numerical measures of human feelings. *Proceedings of the National Academy of Sciences*, *119*(42), e2210412119. https://doi.org/10.1073/pnas.2210412119

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*(3), 527–535. https://doi.org/10.1037/0033-2909.112.3.527

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kourany, J. (2020). Might scientific ignorance be virtuous? The case of cognitive differences research. In J. Kourany & M. Carrier (Eds.), *Science and the production of ignorance: When the quest for knowledge Is thwarted* (pp. 123–143). MIT Press. https://doi.org/10.7551/mitpress/12146.003.0018

Maas, H. L. J., van der, K.-J. K., Marsman, M., & Stevenson, C. E. (2017). Network Models for Cognitive Development and Intelligence. *Journal of Intelligence*, *5*(2), 16. https://doi.org/10.3390/jintelligence5020016

Mari, L. P. (2005). Models of the measurement process. In P. Sydenham & R. Thorn (Eds.), *Handbook of measuring system design*. John Wiley & Sons, Ltd. https://doi.org/10.1002/0471497398.mm066

Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*. Routledge/Taylor & Francis Group.

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 University Press. https://doi.org/10.1017/CBO9780511490040

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

Pearl, J. (2009). *Causality* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511803161

Pessoa, L. (2023). The entangled brain. *Journal of Cognitive Neuroscience*, *35*(3), 349–360. https://doi.org/10.1162/jocn_a_01908

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

Poldrack, R. A., & Yarkoni, T. (2016). From brain maps to Cognitive Ontologies: Informatics and the search for mental structure. *Annual Review of Psychology*, *67*(1), 587–612. https://doi.org/10.1146/annurev-psych-122414-033729

Rohde, T. E., & Thompson, L. (2007). Predicting academic achievement with cognitive ability. *Intelligence*, *35*(1), 83–92. https://doi.org/10.1016/j.intell.2006.05.004

Romeijn, J.-W., & Williamson, J. (2018). Intervention and identifiability in latent variable modelling. *Minds and Machines*, *28*(2), 243–264. https://doi.org/10.1007/s11023-018-9460-y

Scheel, A. M., Tiokhin, L., Isager, P. M., & Lakens, D. (2021). Why hypothesis testers should spend less time testing hypotheses. *Perspectives on Psychological Science*, *16*(4), 744–755. https://doi.org/10.1177/1745691620966795

Serpico, D. (2021). The cyclical return of the IQ controversy: Revisiting the lessons of the resolution on genetics, race and intelligence. *Journal of the History of Biology*, *54*(2), 199–228. https://doi.org/10.1007/s10739-021-09637-6

Sherry, D. (2011). Thermoscopes, thermometers, and the foundations of measurement. *Studies in History & Philosophy of Science Part A*, *42*(4), 509–524. https://doi.org/10.1016/j.shpsa.2011.07.001

Spirtes, P., Glymour, C. N., & Scheines, R. (2000). *Causation, prediction, and search*. MIT Press.

Suppes, P., David, M. K., Duncan Luce, R., & Tversky, A. (1989). *Foundations of measurement, vol. 2: Geometrical, threshold, and probabilistic representations*. Academic Press.

Tal, E. (2017). A model-based epistemology of measurement. In N. Mößner & A. Nordmann (Eds.), *Reasoning in measurement* (pp. 233–253). Routledge.

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. (2019). Conjoint Measurement Undone. *Theory & Psychology*, *29*(1), 100–128. https://doi.org/10.1177/0959354318788729

Uchiyama, R., Spicer, R., & Muthukrishna, M. (2021). Cultural Evolution of Genetic Heritability. *The Behavioral and Brain Sciences*, *45*, e152. https://doi.org/10.1017/S0140525X21000893

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

van Bork, R., Epskamp, S., Rhemtulla, M., Borsboom, D., & van der Maas, H. L. J. (2017). What is the p-factor of psychopathology? Some risks of general factor modeling. *Theory & Psychology*, *27*(6), 759–773. https://doi.org/10.1177/0959354317737185

van Bork, R., Rhemtulla, M., Sijtsma, K., & Borsboom, D. (2022). A causal theory of error scores. *Psychological Methods*. https://doi.org/10.1037/met0000521

van Bork, R., Wijsen, L. D., & Rhemtulla, M. (2017). Toward a causal interpretation of the common factor model. *Disputatio*, *9*(47), 581–601. https://doi.org/10.1515/disp-2017-0019

van der Maas, H. L. J., Dolan, C. V., Grasman, R. P. P. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. J. (2006). A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. *Psychological Review*, *113*(4), 842–861. https://doi.org/10.1037/0033-295X.113.4.842

VanderWeele, T. J. (2022). Constructed measures and causal inference: Towards a new model of measurement for psychosocial constructs. *Epidemiology*, *33*(1), 141–151. https://doi.org/10.1097/EDE.0000000000001434

Weidman, A. C., Steckler, C. M., & Tracy, J. L. (2017). The jingle and jangle of emotion assessment: Imprecise measurement, casual scale usage, and conceptual fuzziness in emotion research. *Emotion*, *17*(2), 267–295. https://doi.org/10.1037/emo0000226

Wilshire, C. E., Ward, T., & Clack, S. (2021). Symptom descriptions in psychopathology: How well are they working for us? *Clinical Psychological Science*, *9*(3), 323–339. https://doi.org/10.1177/2167702620969215

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford University Press.

Woodward, J. (2015). Methodology, ontology, and interventionism. *Synthese*, *192*(11), 3577–3599. https://doi.org/10.1007/s11229-014-0479-1

Yuste, R. (2015). From the neuron doctrine to neural networks. *Nature Reviews Neuroscience*, *16*(8), 487–497. https://doi.org/10.1038/nrn3962