Heating up the measurement debate: What psychologists can learn from the history of physics

Laura Bringmann
Markus Eronen

Forthcoming in *Theory & Psychology*

Abstract: Discussions of psychological measurement are largely disconnected from issues of measurement in the natural sciences. We show that there are interesting parallels and connections between the two, by focusing on a real and detailed example (temperature) from the history of science. More specifically, our novel approach is to study the issue of validity based on the history of measurement in physics, which will lead to three concrete points that are relevant for the validity debate in psychology. First of all, studying the causal mechanisms underlying the measurements can be crucial for evaluating whether the measurements are valid. Secondly, psychologists would benefit from focusing more on the robustness of measurements. Finally, we argue that it is possible to make good science based on (relatively) bad measurements, and that the explanatory success of science can contribute to justifying the validity of measurements.

Keywords: measurement; temperature; validity; physics; psychology; theory

1. Introduction

Current theories of psychological measurement are largely disconnected from discussions of measurement in the natural sciences. There are extensive debates on measurement in both domains, but attempts to bring them together are rare.[1] In most text books, manuals, or monographs on psychological measurement (e.g., AERA, APA, & NCME, 2014; Borsboom 2005; Kline 2000; McDonald 1999), physical measurement appears (if at all) only in the form of simplified standard examples, such as weight and length, which are not analyzed in any serious detail, and are often used just as contrast cases to psychological measurement.

Indeed, there are substantial differences between psychological and physical measurement: human subjects are capable of learning and thus may react very differently at different time points, psychological measurements usually are sum scores made up of item responses, the results may have social and ethical implications, there is a difference between intra- and inter-individual measurements, and so on (Gigerenzer, 1987; McDonald, 1999; Messick, 1989). However, our approach in this paper is to focus on the similarities, and to look for aspects and episodes in the history of physical measurement that are relevant for psychological measurement, more specifically for the debate on validity.

Validity is arguably the most fundamental and controversial issue in psychological measurement (Lissitz, 2009). In the latest edition of the *Standards for Educational and Psychological Testing*, validity is the first topic discussed, and is characterized as "the most fundamental consideration in developing tests and evaluating tests" (AERA et al., 2014, p.

[1] Some notable exceptions are Humphry (2011, 2013), Michell (1999) and Rasch (1980),.

11). According to the classic definition, validity refers to the extent to which the test or instrument measures what it is intended to measure (Kline, 2000, p. 17; McDonald, 1999, p. 197), but in contemporary validity literature, accounts of validity differ greatly, and there is no agreement even on how validity should be defined (see Newton & Shaw (2013) for a state-of-the-art overview).

Some of the most prominent approaches to validity are Messick's (1989) unified treatment of validity (and its various refinements), where the focus is on the adequacy of inferences that psychologists make based on test scores; Kane's (2001, 2006, 2013) argument-based approach, where validation consists in giving evidence-backed arguments for interpretations of test scores; and the causal approach, where the idea is that in a valid measurement the thing to be measured should cause the measurement outcome (Borsboom, 2005; Borsboom, Mellenbergh & van Heerden, 2004; Markus and Borsboom, 2012).[2] In this paper, we take as the starting point the core idea that validity concerns the extent to which the instrument or test measures what it is intended to measure, but our conclusions have relevance independently of how validity is defined.[3]

Natural scientists do not use the same terminology as psychologists or psychometricians, and do not talk of the validity of measurements, but this does not mean that issues related to validity do not arise in physics – as we will show in this paper, in the history of physics it has often been unclear whether the instruments are measuring what they are intended to measure. In this paper, we will focus on a real and detailed example (temperature) from the history of science, and establish connections and parallels between physical and psychological

[2] Interestingly, Hood (2009) argues that the causal approach is in fact compatible with Messick's (1989) approach to validity, and Newton & Shaw (2013) argue that Kane's argument-based account is compatible with the causal account. Thus, it may be that the different approaches are compatible and just focus on different aspects of validity.

[3] If validity is defined in a different sense, so that our arguments do not strictly speaking concern validity, they are still relevant for psychological measurement in general.

measurement.[4] More specifically, our novel contribution is to show that looking at the history of measurement in physics can lead to new insights and viewpoints for the validity debate in psychology.

We have chosen temperature as our case study because temperature measurement has a long and rich history that is well documented, and has been analyzed in detail by historians and philosophers of science (e.g., Chang, 2004; Sherry, 2011). Furthermore, temperature is a representative example of a physical attribute that can be measured in various ways, and that is easily understandable without any background in physics. As we will show, there are surprising parallels between temperature measurement in the first half of the 19th century and the current situation in psychological research practice. In that period, the focus was on making temperature measurements increasingly precise, consistent, and mutually compatible, without engaging in theoretical work on the nature of temperature. In a similar way, current practice in psychological measurement focuses mainly on criteria such as reliability, generalizability, and correlation with other measures, and far less on theories concerning the psychological attributes measured, or the causal processes underlying the measurements (Borsboom, 2005; Markus & Borsboom, 2012, Hubley, Zhu, Sasaki, & Gadermann, 2013).[5]

As we will show, in temperature measurement this atheoretical approach was insufficient, and substantial scientific progress was made only when the measurements were linked to theory. At a very general level, this supports the views in psychology that emphasize the crucial importance of theory for measurement and validity. However, our main point is to analyze three more concrete conclusions that can be drawn from the physical case, and that are

[4] In this paper, we understand psychological measurement to cover all kinds of measurements done in the various fields of psychology, ranging from intelligence tests to measurements of the capacity of short-term memory.

[5] It should be noted that this concerns research and measurement *practice* in psychology. Theoretically oriented psychologists have discussed the importance of theory throughout the 20th century and up to this day.

relevant for the validity debate in psychology. First of all, studying the causal mechanisms underlying the measurements can be crucial for evaluating whether the measurements are valid. Secondly, psychologists would benefit from focusing more on the robustness of measurements. Robustness refers here to the idea that if there are several independent ways of measuring something, this increases our confidence in the measurements.[6] Finally, we argue that it is possible to make good science based on (relatively) bad measurements, and that the explanatory success of science can contribute to justifying the validity of measurements.

As an important terminological remark, the expressions 'mechanism' and 'causal' in this paper should be understood very broadly. By 'mechanism' we mean roughly a set of components that are organized together to perform a function (Bechtel, 2008). This covers not only physical mechanisms, but also cognitive and biological mechanisms that need not be deterministic. Similarly, 'causal' and 'causation' should be understood here in terms of difference-making and potential manipulation and control (Woodward, 2003), and not in terms of exclusively deterministic or physical causation. These broad conceptions of causation and mechanism are compatible with the possibility that the mind or the brain is fundamentally probabilistic (cf. Gigerenzer, 1987).

The structure of this paper is as follows. In the next section, we will briefly sketch the relevant cases from the history of temperature measurement. In section 3, we will relate this to the current situation in psychological measurement, and discuss in detail three insights from the history of temperature measurement that are relevant for the validity debate in psychology. In

[6] The term 'robustness' is ambiguous, and can refer to different things in different contexts. For example, Markus & Borsboom (2012) use it to characterize the stability of causal relationships, and in statistics, it refers to measures that are resistant to deviations and errors. We use the term in order to make a connection to the long tradition in philosophy of science (going back at least to Wimsatt (1981)), where robustness has been discussed in the same sense as in this paper.

section 4, we discuss open issues and briefly return to the general topic of measurement in psychology and physics.

2. A brief history of temperature measurement

In this section, we will briefly go through some key episodes in the history of temperature measurement. The main focus will be on the atheoretical approach to measurement that reached its high point in the work of Henri Victor Regnault. As we will argue, this approach resulted in very precise and comparable temperature measurements, but fell short for various reasons. Most importantly, it did not result in increased understanding of what temperature is, and it did not help in assessing what happens in new circumstances when the validity of measurements is unclear. Furthermore, we point out that the high degree of precision, consistency and comparability that Regnault was aiming at is not even necessary for making valid measurements, and that theoretical progress can provide indirect evidence for the validity of measurements.

Let us start with early days of temperature measurement (see Barnett, 1956 for historical details; the following overview is mostly based on Chang, 2004, pp. 39-56). In the 16th century, researchers such as Galileo started making attempts to develop instruments (thermoscopes) to be able to measure phenomena of heat and cold. Based on subjective sensations of warm and cold, it was discovered early on that liquids (and air) tend to expand when they are heated. Thus, a liquid in a closed glass tube (or any closed vessel) will expand as it gets warmer, and contract as it gets colder. This principle was the starting point for measuring heat and cold. What this means is that the measurement of temperature was not originally based on physical theory, but started from subjective experiences and a simple empirical regularity.

As Chang (2004, pp. 51-52, 159) has pointed out, the improvement of the precision and consistency of temperature measurements proceeded iteratively without much influence from theoretical developments. The simple thermoscopes described in the previous paragraph made it possible to find phenomena that are relatively constant in temperature (such as boiling), and these could be used as fixed points for measurements. Based on this, it was possible to divide the interval between two fixed points into units, resulting in a numerical temperature scale, which allowed for more precise measurements. These numerical thermometers could then be made better and better in terms of various empirical criteria: They could be made more *precise* in the sense that they produce more fine-grained readings, more *consistent* (or reliable) in the sense that they produce the same result in the same circumstances, more *comparable* in the sense that any two particular thermometers of the same type function in the same way, and more *robust* in the sense that different types of thermometers give the same results. In this way, measurements could be improved to a high degree, independently of theoretical developments (see also Choppin, 1985).

The culmination of this approach of improving thermometers based on empirical criteria was the work of the French scientist Henri Victor Regnault (1810-1878). Regnault shunned all theoretical speculation about the nature of temperature and emphasized the importance of rigorous testing with a minimal amount of assumptions (Barnett, 1956, pp. 333-341; Chang 2004, pp. 74-84). Thus, Regnault's approach was anti-theoretical to the extreme. He collected a vast amount of data based on meticulously precise measurements, and used different constructions of instruments to make sure the results were robust (Chang, 2004, p. 175). In the end, Regnault successfully constructed highly precise and comparable gas thermometers, the measurements of which differed from each other only by less than 0.1% (i.e., if one thermometer recorded a temperature of 70 ºC, the measurements of the same conditions by the other thermometers fell within the range 69.93 - 70.07 ºC; Chang, 2004, p. 81).

However, even though Regnault was able to make such extremely precise and consistent measurements, his approach had some fundamental shortcomings. First of all, although an atheoretical approach can guarantee that measurements are consistent and comparable in controlled conditions, such an approach falls short when the conditions are new or unknown. For example, in the 18th century, the best thermometers available were mercury thermometers (invented by Fahrenheit).[7] They had been extensively tested and used only in conditions that naturally occur or that are easy to produce in a laboratory, but it was unclear whether they would continue to provide valid measurements in other circumstances, such as extreme heat or cold. In fact, they did not, as is nicely illustrated by the following story (described in Chang, 2004, pp. 105-118).

In 1733, the Russian scientist Johann Georg Gmelin set out to explore the eastern stretches of Siberia, and on his journey experienced freezing conditions of unexpected harshness. The mercury thermometer that Gmelin was using indicated a temperature of -120° Fahrenheit (-84.4° C). Gmelin was happy to accept this reading as roughly accurate, as indeed it had been very cold, but others were skeptical. Nothing even close to that temperature had ever been recorded on earth. It seemed more likely that the thermometer was no longer providing valid measurements of temperature. This initiated a heated scientific debate, and a new research project: Although many of the properties of mercury were well understood, its freezing point was unknown, which various scientists now set out to discover. Only decades later it was established that mercury freezes around -40° C. Like most substances, mercury becomes much denser when frozen, resulting in lower levels of mercury in the thermometer. Thus, the

[7] This episode took place before Regnault's time, but we describe it here because it gives a vivid illustration of the limits of focusing just on precision and reliability, and thus also the limits of Regnault's approach.

mercury of Gmelin's thermometer had frozen, and the temperature had been far less severe than -84.4° C.

The crucial point here is that, in contrast to what Gmelin thought, the comparability and consistency that had been established for mercury thermometers provided no justification for believing in the results, because the conditions were novel and untested. In a similar way, the precision, consistency and comparability of Regnault's thermometers was only established for limited conditions and a limited part of the temperature scale. Furthermore, it was not possible to resolve the issue of whether the measurements continued to be reliable based on just empirical criteria. Any other mercury thermometer would have also frozen in conditions of extreme cold. In order to solve the problem, scientists had to study how the measurement instrument actually works, i.e., the causal mechanism that results in the measurement outcome, and this in turn required theoretical advances (i.e., discovering that a substance like mercury can freeze, and that it will contract when frozen).

A second limitation of the atheoretical approach of Regnault was that this approach did not result in increased understanding of what temperature is, or in new connections with other areas of physics. One concrete implication of this was that the temperature scale itself remained just as arbitrary as it had been before Regnault's efforts: The fixed points of the scale(s) were conventions based on practical considerations, and there was no plausible theoretical definition for what it means for temperature to change by one degree. Only after a connection was made to theory did it become possible to formulate an objective definition for what a change of one degree of temperature means, and to calculate the absolute zero (Chang, 2004, pp. 159-197). This was achieved by Thompson (also known as Lord Kelvin, 1824-1907), who connected temperature to the thermodynamic notions of work and heat. This also allowed making temperature measurements more robust: While Regnault's measurements were all in the end based on simple gas laws, the connection to thermodynamics made it

possible to derive temperature values also from thermodynamic equations. Eventually, other theoretical advances also resulted in new kinds of instruments for measuring temperature, such as resistance thermometers, which are based on the principle that the electrical resistance of some materials increases with rising temperature.

As a third point, in order to achieve valid measurements and scientific progress, the kind of high degree of precision, consistency and comparability that the atheoretical approach aims at is not necessary – to put it simply, it is possible to do good science based on relatively bad measurements. To illustrate this, we can again move back in the history of temperature measurement, and consider an episode that took place before Regnault's time: Joseph Black's (1728-1799) discovery of the theory of latent heat (the heat that a substance can absorb or release without changing in temperature; Sherry, 2011). Black's theory amounted to a great scientific advance that marked the beginning of the science of thermodynamics. Interestingly, the measurement instruments at Black's disposal were mercury thermometers that by Regnault's standards (or contemporary standards) would not have counted as very consistent, precise, or mutually compatible. However, this did not stop Black from theorizing about heat and temperature and testing his hypotheses with temperature measurements. Even with these imperfect measurements, Black was able to quantify the notion of latent heat, and with his theory of latent heat he could provide novel explanations to a broad range of phenomena, including the melting of ice and freezing of water.

This case also illustrates another related point: Theoretical progress can contribute to the validity of measurements retroactively, or in hindsight. The theory of latent heat was built on the assumption that mercury thermometers provide valid (although imprecise) measurements of temperature. The predictions and explanations based on the theory of latent heat were extremely successful. Thus, Black and his contemporaries had good reasons to believe that the original hypothesis concerning the validity of the measurements was correct (Sherry, 2011).

Of course, it was in principle possible that the theory happened to be correct in spite of the temperature measurements being completely invalid, but this would have been almost miraculous: The far more likely explanation was that the measurements were in fact valid, at least in the sense that they were roughly measuring some real quantity (Sherry, 2011). Thus, theoretical progress and success can contribute to indirectly justifying the validity of measurements.

3. Lessons for Measurement in Psychology

The most general moral that we can draw from the above is that theory is crucially important for measurement and validity. While an atheoretical approach, where the aim is to make measurements better on purely empirical standards (such as reliability and invariance), will result in measurements that are consistent and comparable under a limited range of conditions (corresponding to Regnault's achievements), it will not guarantee the validity of measurements or lead to significant scientific progress.

In many ways, the situation in psychological research practice resembles the situation of temperature measurement in the late 18th and early 19th century: The focus is on criteria such as reliability and invariance, and on correlational and purely empirical studies, at the expense of theory-building or theoretical speculation. The standard approach in psychometric modeling is to find statistical models that fit the data, which can be done independently of theoretical assumptions concerning the thing that is measured (Markus & Borsboom 2013, p. 43). Assessments of validity in practice most often amount to evaluating the internal structure of the test, or correlating the results with external variables and seeing whether the correlations are in the right direction (Borsboom, Cramer, Kievit, Scholten, & Franić., 2009; Borsboom et al., 2004, Hubley et al., 2013). Furthermore, just like the temperature scales in

Regnault's time, psychological scales and units lack any clear theoretical foundation, and there is no clear understanding of the nature of the attributes measured (Humphry, 2011).

The temperature story illustrates the limits of such an atheoretical approach: It will result in measurements that are consistent and precise under a limited range of conditions (corresponding to Regnault's achievements), but it will not guarantee the validity of measurements or lead to significant scientific progress. For advances on these fronts, theory is required.

This point as such is not novel – the importance of theory has been widely discussed in the debates on validity in psychology, starting from the classic paper by Cronbach & Meehl (1955) and going on up to this day (e.g., Borsboom, 2005; Embretson, 1983; Embretson & Gorin, 2001; Kane, 2013; Markus & Borsboom, 2012; Messick, 1989; Newton & Shaw 2013). Here we have provided new evidence for the perils of neglecting theory. Furthermore, in the validity literature, the views on the role and importance of theory for establishing or assessing validity vary greatly, and the considerations in the previous section clearly support accounts that place theory at the very core of assessing validity (e.g., Borsboom et al., 2004; Embretson, 1983, 1998; Embretson & Gorin 2001).

However, our main point in this section is to show how the three more specific issues we picked up from the history of temperature measurement are relevant for the validity debate. Our first main point in the previous section was that understanding the causal mechanism underlying the measurement instrument is essential for assessing validity. This was evident in the story of Gmelin's frozen thermometer: In order to determine whether Gmelin's measurements were valid measurements of temperature, scientists had to study what exactly happens in the mechanism of the thermometer in extreme temperatures. Thus, theoretical understanding of the causal mechanism of measurement seems to be crucial for assessing the

validity of measurements, especially in novel circumstances. More generally, in the philosophical literature on measurement, the focus is nowadays on understanding and modeling the measurement process and the (causal) functioning of the measurement instrument (Chang, 2004; Frigerio, Giordani, & Mari, 2010; Kyburg, 1992; Tal, 2013; Trout, 1998).

This suggests that understanding the causal mechanisms underlying measurements should be crucial also for assessing psychological validity, in line with the causal account of validity defended by Borsboom and colleagues (Borsboom, 2005; Borsboom et al., 2004; Borsboom et al., 2009; Markus and Borsboom, 2012). However, we do not agree with these authors that validity *only* concerns the question of whether the attribute to be measured actually exists and causes the variations in the measurement outcomes (e.g., Borsboom et al., 2004, p. 1061). As will become clear below, we believe that there are also many other important aspects to validity.

One obvious problem that arises when the approach of studying the causal mechanisms of measurement is applied psychology is whether we can actually study the relevant mechanisms. The most prominent attempt at this is found in Susan Embretson's groundbreaking work. According to Embretson's (1983, 1998, 2004) process-oriented account of validity, traditional assessments of construct validity (i.e., comparing the test scores to relevant external variables) need to be supplemented with studies of the cognitive processes and strategies that participants use to respond to test items, based on the state-of-the-art cognitive psychology. When this approach is applied in practice, cognitive theory influences both test construction and the measurement models (such as item response theory (IRT) models): The items selected for the test are based on cognitive theory, and the models include parameters representing the cognitive demands of the item (Tatsuoka, 1987, 1990;

Embretson, 1998, 2004).[8] For example, a test for assessing abstract reasoning was created based on processing theory, and the IRT model was combined with a cognitive model, including parameters such as working memory load and perceptual processing (Embretson, 1998).

However, as important as these studies are, including cognitive parameters in measurements models is still very far from describing the causal mechanisms underlying the measurement process. Ideally, there should be models that describe the steps in the causal process that starts with the attribute intended to measure and ends with the measurement outcome. It may be that the reason why this is not generally attempted, or not even seen as a goal, is that the causal mechanisms in psychology are thought to be so complex that figuring out the causally relevant components is practically impossible (Trendler, 2009). We acknowledge that the challenges may seem daunting at present, but do not believe that the situation is hopeless with regard to the future: Great progress has been made in recent decades in discovering mental mechanisms (Bechtel 2008), an issue to which we return below.

The second important insight for validity (and psychological measurement in general) that we draw from the physical case is the principle of robustness. This is a method or principle that pervades all of science and has also many other names: mutual grounding, mutual compatibility, overdetermination, triangulation, diverse testing, and so on (Chang, 1995, 2004; Eronen, 2012; Hacking, 1983; Tal, 2011; Trout, 1998; Wimsatt, 1981, 2007). The basic idea is that if there are several independent ways of achieving the same result, this increases our confidence in the result. This can also be expressed as the following mathematical principle: If there are several (independent) ways of measuring something, the probability that all of them happen to go wrong is a product of the individual probabilities of going wrong,

[8] A similar approach to incorporating cognitive theory into test development is Mislevy's Evidence Centered Design (ECD) (Mislevy, Steinberg, & Almond, 2002).

and this product becomes increasingly tiny as more and more independent ways are added (Wimsatt, 1981).

The idea of independence is crucial for robustness. There is no uncontroversial or widely accepted account of the exact nature of the required independence, but certain key features can be spelled out (Nederbragt, 2012; Stegenga, 2012; Stroebe & Sack, 2014; Wimsatt, 2007). First of all, it is obvious that statistical independence is not what is required: Different ways of measuring temperature will be statistically correlated, even when they are in other important respects independent (e.g., two thermometers based on different physical principles, such as a mercury thermometer and a radiation thermometer). The idea is rather that the different ways of measuring should partly rely on different theoretical assumptions, different physical processes, or different experimental setups. What is necessary is that any problematic or unconfirmed assumptions should not be shared by the different ways (Stegenga, 2012). Different approaches or ways of measuring are fully independent only if they rely on different assumptions and different parts of theory (such as mercury thermometers and radiation thermometers). It is clear that independence is a matter of degree, and not a none-or-all property: Two different mercury thermometers are less independent from each other than a mercury thermometer and a radiation thermometer.

Robustness itself is also a matter of degree, corresponding to the number and independence of the different ways of measuring. Once a high degree of robustness is reached, we can be confident in the measurements, and conversely, if measurements are not robust or just robust to a low degree, we should approach them with healthy skepticism. This principle can be applied to measurements, attributes, properties, and entities, but it is important to keep in mind that it is a fallible epistemic principle, and not a guarantee of truth or reality (see Eronen (2012) and Wimsatt (2007) for more).

An increase in the degree of robustness is evident in the history of temperature measurement, as outlined in the previous section. If several different thermometers give the same reading, it is more likely that the reading is correct than when only one thermometer is used. However, applying multiple instruments based on the same theoretical principle leads to a low degree of robustness, because their independence from each other is very limited. If the theory on which the instruments are based turns out to be false, the fact that multiple instruments were used becomes irrelevant.[9] Also the robustness of Regnault's measurements was limited, because all the thermometers he used were implicitly based on the same gas laws, and no further connection to theory was made. A higher degree of robustness was only reached after theoretical developments, leading eventually to new types of instruments for measuring temperature, relying on different areas of physics, such as radiation thermometers and resistance thermometers.

The idea of robustness (although different terms are used) also has a tradition in psychometrics, going all the way back to the classic paper by Cronbach & Meehl (1955) and the multitrait-multimethod matrix approach of Campbell & Fiske (1959). In contemporary validity theory, the idea comes up in the context of convergent validity, which refers to evidence from other measures that are intended to assess the same or similar attribute (AERA et al., 2014, pp. 16-17). However, in practice, assessing convergent validity usually amounts to calculating correlation coefficients between measures that are expected to be related,

[9] A classic example of this is the bacterial mesosome (Culp, 1994; Wimsatt, 2007, p. 381, note 3). This entity appeared in various experiments studying bacteria and was initially thought to be a new kind of cellular organelle. Because independent research groups using different experimental setups could detect the bacterial mesosomes, the results that indicated their existence seemed to be robust. However, it later turned out all of the different experimental setups were using the same fixation methods for preparing the samples, and the "bacterial mesosomes" were merely artifacts of the preparation methods. Thus, the various experimental setups were in a crucially important respect not independent from each other, and the robustness of the results was only apparent.

possibly supplemented with factor analysis or principal component analysis (see, e.g., Clapham, 2004; Duckworth & Kern, 2011). No special attention is paid to the independence of measures, or to deriving the result from theory. Thus, convergent validity can be seen as a weak form of robustness. Furthermore, in psychometrics convergent validity is usually just briefly mentioned as one possible source of evidence for validity, while in the natural sciences robustness is central to the validity of measurements (Chang 1995, 2004; Wimsatt, 2007).

It is clear that the degree of robustness of the measures and constructs in contemporary psychology varies greatly. For example, it could be argued that intelligence measurements (IQ scores) are robust to a low degree, because although the results of different intelligence tests are highly correlated, intelligence scores are not based on any widely accepted theory, and the independence of the various tests can be questioned (van der Maas et al., 2006). An example of a domain where psychological measurements have a higher degree of robustness would be short-term memory. The capacity of short-term memory can be measured in a broad range of different types of (independent) experimental setups: imposing an information overload, preventing long-term memory access, examining performance discontinuities in memory tasks, mathematically modelling memory performance, and so on (Cowan, 2000; Jonides et al., 2008). In any case, we believe that the debate on validity would benefit from a closer analysis of robustness, and more specifically the independence of measurements.

The third main point for validity that we draw from the history of physics can be summarized as follows: Relatively bad measurements can result in good science, and scientific progress can justify the validity of measurements in hindsight. This was discussed in the last part of the previous section. Instead of waiting for better measurement techniques, theoretically-oriented scientists such as Black made the working hypothesis that the imperfect measurement instruments (mercury thermometers) at their disposal were consistent and valid enough, and formulated theories and explanations based on that working hypothesis (see also Choppin,

1985). Those theories turned out to be very successful. This success gave the scientists more reason and justification for believing that the working hypothesis was true, and that the measurements were valid in the sense that a real and scientifically important quantity was being measured.[10] Note that this does not involve any vicious circularity in the sense that measurements are validated by theory, and the theory is validated by measurements. The pattern is rather this: What increases confidence in the validity of measurements is the success of the theories that are based on them, and what justifies the success of those theories is their explanatory and predictive power. Testing the latter need not involve the same types of measurements whose validity is in question.

We certainly do not want to claim that this is the only way of establishing the validity of measurements. The point is rather that this is one way of contributing to validity arguments or justifying validity claims. To the best of our knowledge, this has not been explicitly discussed in current validity literature (although interestingly Coleman (1964, pp. 70-73) makes a similar point in the context of measurement in sociology). Consequences are generally regarded as one source of evidence for validity, but this refers to the consequences of test use in practice, and not consequences for theory and science in general (see, e.g., Messick, 1989; Newton & Shaw, 2013).

This point also has implications for Joel Michell's arguments against psychological measurement. Michell (1997, 1999, 2000, 2013) has argued that psychologists are treating the attributes they are measuring as quantitative, without having even attempted to show that they

[10] There are also numerous cases in the history of science where the working hypothesis was not confirmed, and the measurements turned out to have been invalid. Consider phlogiston: scientists in the 17th and the 18th century assumed that they were measuring quantities of phlogiston in combustible things. However, explanations and predictions based on the phlogiston theory were fundamentally problematic, and eventually the theory was replaced by oxygen-based theories of combustion. Thus, the validity of phlogiston measurements was disconfirmed by later developments in science.

fulfil the requirements for being quantitative in structure (such as additivity, i.e., that there is a meaningful way of adding up quantities of the attribute). Thus, psychological measurement currently lacks any foundation, and as long as it has not been shown that psychological attributes are in fact quantitative, psychometrics is a "pathological" science (see also Trendler, 2009).

However, as Sherry (2011, pp. 515-517) has pointed out, the history of thermometry suggests that this shortcoming may be less devastating (or "pathological") for psychology than Michell thinks. Black and his contemporaries did not have any conclusive arguments for the quantitative nature of temperature either, and they had no conception of the actual nature of heat and temperature (which were only discovered in the late 19th century; see also Choppin, 1985). Instead, they made the working hypothesis that temperature is measurable and that mercury thermometers provide roughly valid measurements of it, and built their theories based on this assumption. Considering the success of those theories, in retrospect it is clear that making that hypothesis was a crucially important and justified move. Thus, it is plausible that psychologists can also make a similar working hypothesis, which will then be confirmed or disconfirmed by later developments in science (see also Humphry 2011).

Michell could respond that while this strategy works in physics, there is no reason to expect it to work in psychology. In Black's case, the theories based on temperature measurements were very successful, and were soon broadly accepted, but psychology has so far failed to produce theories of significant scope or explanatory power, and current psychological theories are not rich and detailed enough to provide serious tests for the hypothesis that psychological measurements are valid (see, e.g., Michell, 2004). Thus, perhaps we cannot expect in psychology the kind of progress that led to the vindication of temperature measurements.

In our view, this is a question that can only be resolved by the eventual development and progress of psychology as a science, and in this regard we are far less skeptical than Michell. At the moment, no overarching theories of the kind developed by Black, Thomson, or Maxwell in the case of temperature are foreseeable in psychology, but this should not be seen as discouraging: Theories in psychology and the life sciences in general tend to be more local than in physics (Bechtel 2008; Bechtel & Richardson, 1993; Kyngdon, 2013). For example, there is no (and likely never will be) single overarching theory of biology, but a broad range of theories concerning natural selection, gene expression, development, ecology, and so on. In a similar way, instead of one unified theory of human psychology, there will be increasingly precise theories or models of perception, language learning, problem solving, and so on.

In fact, it may be that theoretical development in psychology is hampered by the implicit assumption that theories are simply the better the more general they are. For example, based on the mirror neuron mechanism, psychologists have proposed far-reaching theories of social cognition (Gallese, Rochat, Cossu, & Sinigaglia, 2009). It may be better to focus first on local mechanisms and their specific and limited roles and functions, for example the role of mirror neurons in perception of goal-directed behavior (Spaulding, 2013), and to make these local theories and explanations as elaborate and plausible as possible.

Indeed, in many areas of psychological measurement, such as memory, we already find relatively well-developed local theories, for example in the case of short-term memory mentioned above. There are theories concerning the interplay of short-term memory, long-term memory, focus of attention, and perception (Cowan, 2000; Jonides et al. 2008). These theories predict and explain what happens, for example, when subjects are asked to report the elements in a visual array that is presented in a flash: Very roughly, their focus of attention has a limited capacity, and thus subjects only report the elements they attended to (Cowan

2000). Other examples of promising and local psychological theories are provided by Kyngdon (2013).

4. Concluding remarks

In this paper, we have discussed the history of the measurement of temperature and its methodological relevance for psychology, particularly the debate on validity. We started by going through some important episodes in the history of temperature measurement, and pointed out that there is a parallel between the hyper-empirical approach of Regnault in the 19th century and the atheoretical attitude that is still common in psychological research practice. We also argued that this approach was in the end insufficient in temperature measurement, and that it will likely be insufficient in psychology as well. In short, it does not lead to increased understanding of the phenomena or attributes measured, does not lead to important scientific advances, and the high reliability and consistency that it strives for is not even necessary for valid measurements.

In section 3, we then looked at more concrete ways in which the validity of psychological measurements could be improved. Interestingly, all of these three ways are closely related to theory. Our first point was that assessing measurements in novel situations requires theoretical understanding of the causal mechanism underlying the measurement process. In the case of robustness, determining the degree of robustness depends on theories and models about the experimental setups, measuring instruments, and the things being measured. Our last point explicitly concerned theories: The validity of measurements can be indirectly justified or established based on the development of successful theories that build on the measurements. Thus, this article can be seen as continuing the long tradition of emphasizing the importance of theory for measurement and validity.

Although we have argued that there are parallels between physical and psychological measurement, we do not want to deny that there are also substantial differences, as we acknowledged in the introduction. However, we believe that the differences are a matter of degree, and not as categorical as is often supposed. For example, although properties such as length or weight can be measured in a relatively direct and straightforward way, the same does not apply to phenomena such as the weak nuclear force or the background radiation of the universe. Such phenomena (which includes most phenomena studied in contemporary physics) can be measured only indirectly, and have no straightforward operationalizations (Kyburg 1984). To return to the example of temperature, it is worth mentioning that even nowadays temperature measurement faces considerable practical and conceptual problems and is far from trivial. For example, a monograph on temperature measurement by Quinn (1990) uses hundreds of pages to discuss various issues and complications in measuring temperature.

In our view, the main differences between physical attributes and psychological attributes are that (1) most psychological attributes have not been embedded into any successful and widely accepted theory (see also Sijtsma, 2012), and relatedly, (2) there is no solid theoretical foundation for the units, ratios and scales for psychological attributes (see also Humphry, 2011, 2013). However, as we have shown, the situation in early temperature measurement was not much better. Thus, we do not think that these differences rule out the possibility of measuring psychological attributes; rather, they emphasize the importance of developing better psychological theories.

In sum, we believe that the existing discussions of this topic have focused too little on the similarities physical and psychological measurement, and we do not think that the differences are so fundamental that they prevent drawing interesting parallels and looking at physical sciences for insights. More generally, we believe that the methodology and history of physical

measurement can be valuable to psychologists, as we hope to have shown in this paper in the context of the validity debate. We do not claim that psychology will necessarily develop in the same way as physics has developed, but rather that psychologists should not think that the history and theory of measurement in physics and other natural sciences is irrelevant for psychology. In conclusion, we hope that this paper can contribute to opening new pathways for studying psychological measurement.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for Educational and Psychological Testing*. Washington, DC: Author.

Barnett, M. K. (1956). The development of thermometry and the temperature concept. *Osiris, 12,* 269-341.

Bechtel, W. C. (2008). *Mental mechanisms.* London: Routledge.

Bechtel, W., & Richardson, R. C. (1993). *Discovering complexity: Decomposition and localization as strategies in scientific research.* Princeton: Princeton University Press.

Borsboom, D. (2005). *Measuring the Mind.* Cambridge: Cambridge University Press.

Borsboom, D., Cramer, A. O. J., Kievit, R.A., Scholten, A. Z., & Franić, S. (2009). The End of Construct Validity. In R. W. Lissitz (Ed.) *The concept of validity: Revisions, new directions, and applications* (pp. 135-172). Charlotte, NC: Information Age Publishing.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111,* 1061-1071.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81-105.

Calcott, B. (2011). Wimsatt and the robustness family: Review of Wimsatt's re-engineering philosophy for limited beings. *Biology & Philosophy*, *26*, 281-293.

Chang, H. (1995). Circularity and reliability in measurement. *Perspectives on Science*, *3*, 153-172.

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

Choppin, B. (1985). Lessons for psychometrics from thermometry. *Evaluation in Education*, *9*, 9-12.

Clapham, M. (2004). The convergent validity of the Torrance tests of creative thinking and creativity interest inventories. *Educational and Psychological Measurement*, *64*, 828-841.

Coleman, J. S. (1964). *Introduction to Mathematical Sociology*. New York: Free Press.

Cowan, N. (2000). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*, 87–185.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281.

Culp, S. (1994). Defending Robustness: The Bacterial Mesosome as a Test Case. *PSA 1994: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, *1*, 46-57.

Duckworth, A. L., & Kern, M. L. (2011). A meta-analysis of the convergent validity of self-control measures. *Journal of Research in Personality*, *45*, 259-268.

Embretson, S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*, 179-197.

Embretson, S. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3*, 380-396.

Embretson, S., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement, 38*, 343-368.

Embretson, S. (2004). FOCUS ARTICLE: The second century of ability testing: some predictions and speculations. *Measurement: Interdisciplinary Research and Perspectives, 2*, 1-32,

Eronen, M. I. (2012). Pluralistic physicalism and the causal exclusion argument. *European Journal for Philosophy of Science, 2*, 219-232.

Frigerio, A., Giordani, A., & Mari, L. (2010). Outline of a general model of measurement. *Synthese, 175*, 123-149.

Gallese, V., Rochat, M., Cossu, G., & Sinigaglia, C. (2009). Motor cognition and its role in the phylogeny and ontogeny of action understanding. *Developmental Psychology, 45*, 103-113.

Gigerenzer, G. (1987). Probabilistic thinking and the fight against subjectivity. In L. Krüger, G. Gigerenzer, & M.S. Morgan (Eds.), *The probabilistic revolution – Volume 2: Ideas in the sciences* (pp. 11-33). Cambridge, MA: MIT Press.

Hacking, Ian (1983). *Representing and intervening. Introductory topics in the philosophy of natural science.* New York: Cambridge University Press.

Hood, S.B. (2009). Validity in psychological testing and scientific realism. *Theory & Psychology, 19*, 451-473.

Hubley, A. M., Zhu, S. M., Sasaki, A., & Gadermann, A. M. (2013). Synthesis of validation practices in two assessment journals: Psychological Assessment and the European Journal of Psychological Assessment. In B. D. Zumbo & E. K. H. Chan (Eds.) *Validity and Validation in Social, Behavioral, and Health Sciences* (pp. 193 – 213). Dordrecht: Springer.

Humphry, S. M. (2011). The role of the unit in physics and psychometrics. *Measurement: Interdisciplinary Research and Perspectives, 9*, 1-24.

Humphry, S. M. (2013). Understanding measurement in light of its origins. *Frontiers in Psychology, 4*. doi: 10.3389/fpsyg.2013.00113

Jonides, J., Lewis, R. L., Nee, D. E., Lustig, C. A., Berman, M. G., & Moore, K. S. (2008). The mind and brain of short-term memory. *Annual Review of Psychology, 59*, 193.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*, 319-342.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.; pp. 17–64). Westport, CT: American Council on Education/Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1-73.

Kline, P. (2000). *Handbook of psychological testing*. London: Routledge.

Kyburg, H. (1984). *Theory and Measurement*. Cambridge: Cambridge University Press.

Kyburg, H. (1992). Measuring errors of measurement. In C. W. Savage & P. Ehrlich (Eds.) *Philosophical and foundational issues in measurement theory* (pp. 75-91). Hillsdale, NJ: Lawrence Erlbaum.

Kyngdon, A. (2013). Descriptive theories of behaviour may allow for the scientific measurement of psychological attributes. *Theory & Psychology, 23*, 227–250.

Lissitz, R. W. (Ed.) (2009). *The concept of validity: Revisions, new directions, and applications*. Charlotte, NC: Information Age Publishing.

Markus, K., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation and meaning*. London: Routledge.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

Meehl, P. (1983). Consistency tests in estimating the completeness of the fossil record: A neo-Popperian approach to statistical paleontology. In J. Earman (Ed.), *Minnesota studies in the philosophy of science: Vol. X, Testing scientific theories* (pp. 413-473). Minneapolis: University of Minnesota Press.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). Washington, DC: American Council on Education and National Council on Measurement in Education.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology, 88*, 355-383.

Michell, J. (1999). *Measurement in psychology*. Cambridge, UK: Cambridge University Press.

Michell, J. (2000). Normal science, pathological science and psychometrics. *Theory & Psychology, 10*, 639-667.

Michell, J. (2004). Item response models, pathological science and the shape of error: Reply to Borsboom and Mellenbergh. *Theory & Psychology, 14*, 121-129.

Michell, J. (2013). Constructs, inferences, and mental measurement. *New Ideas in Psychology*, *31*, 13-21.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*, 3-67.

Nederbragt, H. (2012). Multiple derivability and the reliability and stabilization of theories. In Soler, L., Trizio, E., Nickles, T., & Wimsatt, W. C. (Eds.). *Characterizing the Robustness of Science: After the Practice Turn in the Philosophy of Science* (pp. 121-145). Dordrecht: Springer.

Newton, P. E., & Shaw, S. D. (2013). *Validity in Educational & Psychological Assessment*. London: Sage.

Quinn, T. J. (1990). *Monographs in physical measurement: Temperature*. London: Academic Press.

Rasch, G. (1980). *Some probabilistic models for intelligence and attainment tests*. Chicago: The University of Chicago Press.

Sherry, D. (2011). Thermoscopes, thermometers, and the foundations of measurement. *Studies in History and Philosophy of Science Part A, 42*, 509-524.

Sijtsma, K. (2012). Psychological measurement between physics and statistics. *Theory & Psychology, 22*, 786-809.

Spaulding, S. (2013). Mirror neurons and social cognition. *Mind & Language, 28*, 233-257.

Stegenga, J. (2012). *Rerum Concordia Discors*: Robustness and discordant multimodal evidence. In Soler, L., Trizio, E., Nickles, T., & Wimsatt, W. C. (Eds.). *Characterizing the Robustness of Science: After the Practice Turn in the Philosophy of Science* (pp. 207-226). Dordrecht: Springer.

Stroebe, W., & Sack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science, 9*, 59-71.

Tal, E. (2011). How accurate is the standard second? *Philosophy of Science, 78*, 1082-196.

Tal, E. (2013). Old and new problems in philosophy of measurement. *Philosophy Compass, 8*, 1159–1173.

Tatsuoka, K. K. (1987). Validation of cognitive sensitivity for item response curves. *Journal of Educational Measurement, 24*, 233-245.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold & M. G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 453-488). Hillsdale, NJ: Erlbaum.

Trendler, G. (2009). Measurement theory, psychology and the revolution that cannot happen. *Theory & Psychology, 19*, 579-599.

Trout, J. D. (1998). *Measuring the intentional world*. Oxford: Oxford University Press.

Van Der Maas, H. L., Dolan, C. V., Grasman, R. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. (2006). A dynamical model of general intelligence: the positive manifold of intelligence by mutualism. *Psychological Review, 113*, 842-861.

Wimsatt, W. C. (1981). Robustness, reliability, and overdetermination. In M. Brewer & B. Collins (Eds.), *Scientific inquiry and the social sciences* (pp. 124–163). San Fransisco: Jossey-Bass.

Wimsatt, W. C. (2007). *Re-Engineering Philosophy for Limited Beings. Piecewise Approximations to Reality*. Cambridge, MA: Harvard University.

Woodward, J. (2003). *Making Things Happen. A Theory of Causal Explanation*. Oxford: Oxford University Press.