Theory & Psychology

Article

Conjoint measurement undone

Theory & Psychology
2019, Vol. 29(1) 100–128

DOI: 10.1177/0959354318788729
journals.sagepub.com/home/tap
SAGE

Günter Trendler
Worms, Germany

Abstract

According to classical measurement theory, fundamental measurement necessarily requires the operation of concatenation qua physical addition. Quantities which do not allow this operation are measurable only indirectly by means of derived measurement. Since only extensive quantities sustain the operation of physical addition, measurement in psychology has been considered problematic. In contrast, the theory of conjoint measurement, as developed in representational measurement theory, proposes that the operation of ordering is sufficient for establishing fundamental measurement. The validity of this view is questioned. The misconception about the advantages of conjoint measurement, it is argued, results from the failure to notice that magnitudes of derived quantities cannot be determined directly, i.e., without the help of associated quantitative indicators. This takes away the advantages conjoint measurement has over derived measurement, making it practically useless.

Keywords

derived measurement, replication crisis, psychometrics, quantity objection, Rasch model

Representational measurement theory (Luce & Suppes, 2002) is arguably one of the most influential theories of measurement. Its beginning can be traced back to Suppes' (1951) and Scott and Suppes' (1958) reformulation of Hölder's (1901) axiomatic approach in terms of Tarski's (1954) theory of models. The first systematic exposition was delivered by Suppes and Zinnes (1963) and the definite version is presented in the first volume of *Foundations of Measurement* (Krantz, Luce, Suppes, & Tversky, 1971). The great merit of the axiomatic approach is to have answered theoretical questions of measurement in a consistent mathematical framework. Thus, "the answers to questions of measurement have the same unambiguous status as the answers to mathematical questions posed in other fields of science" (Suppes & Zinnes, 1963, p. 3). Whatever one may think of the philosophical theory underlying the representational approach, it has, Michell

Corresponding author:
Günter Trendler, Robert-Schuman-Str. 15, 67549 Worms, Germany.
Email: guenter@trendler.org

(2017) notes, "furnished us with mathematical results invaluable for an adequate understanding of measurement" (p. 423).

Undoubtedly, the most original contribution of abstract measurement theory is the development of conjoint measurement theory as a new type of fundamental measurement (Luce & Tukey, 1964). Conjoint measurement is considered not only a revolution in measurement theory (see Michell, 1999, Chapter 8), but also "one of the most important developments in scientific psychology" (Cliff, 1992, p. 186), since it holds out the prospect that psychological attributes will finally be measured on interval or ratio scales. Thus, psychology is transformed into a quantitative science on equal footing with physics. There is legitimacy in saying that conjoint measurement was especially developed for the purpose of making psychological attributes measurable.

In retrospect, however, the faith put in this new theory may have been premature. Already, Cliff (1992) referred to abstract measurement theory as the "revolution that never happened" (p. 186), since "its influence on the mainstream of any aspect of quantified psychology has been minimal" (p. 187). More than half a century after the publication of Luce and Tukey's seminal article, the situation is basically unchanged. Of course, one can also argue that the reason for this state of affairs is due to the fact that only a few attempts to apply conjoint measurement have been made (for details on this point see Michell, 1999, pp. 211–216). On the other hand, it must be noted that, for instance, the extensive experimental studies undertaken by Luce (2000) and others in the context of utility theory did not lead to a breakthrough with regard to measurability. Quite the opposite: the unrestricted applicability of the axioms of conjoint measurement was called into question by Luce (2011) himself.[1] In conclusion, as a matter of fact the unsatisfactory situation endures; that is, so far, no interval or ratio scales have been established in psychology, neither by conjoint measurement nor by any other means.

Hence, if one does not want to abstain from using methods of data analysis which rely on the assumption that the measurement problem has been successfully solved (e.g., structural equation modeling, Bollen, 1989; or multiple regression/correlation analysis, Cohen & Cohen, 1975) the question of measurability is to be categorized as urgent. As has been emphasized by Paul Barrett (2008), the consequence of ignoring the issue is scientific stagnation (see also Barrett, 2018). Apparently, the acuteness of the problem is still underestimated. To a substantial extent, the reason for the ignorance is probably due to the fact that most psychologists are not aware of the extent to which stagnation really characterizes psychology as an empirical science. In my view the main characteristic of scientific stagnation is the lack of reproducibility. Obviously, *as long as empirical results are not replicable they cannot be accepted as scientific knowledge*. And without knowledge there is no accumulation of knowledge and in consequence no progress. Unfortunately, in psychology, the topic of reproducibility is usually ignored. However, in a recent large-scale study by the Open Science Collaboration (2015) the reproducibility of psychological science was systematically investigated. The result of the investigation is rather sobering: "After this intensive effort to reproduce a sample of published psychological findings, how many of the effects have we established are true? Zero. And how many of the effects have we established are false? Zero" (p. aac4716-7). I will return to the issue of measurability and reproducibility below.

In previous publications, I have already questioned the utility of conjoint measurement in psychology (Trendler, 2009, 2013). My main argument is that in psychology we are not even able to satisfy the first condition of quantity which demands the identification of equal magnitudes of quantity. What invariably undermines attempts to make progress with regard to measurement are systematic disturbances. The problem can be solved—as it is, for example, in physics—through the construction of experimental apparatus. This method is, in my view, not applicable in psychology to the extent necessary for a successful application of measurement theory. I have called this the Millean quantity objection.

Here I will argue that—irrespective of the validity of the Millean quantity objection—conjoint measurement is a superfluous method for the investigation of measurability of psychological factors. First, I will explore the central role of the so-called standard sequence procedure in representational measurement theory. Relying on this investigation and on the fact that conjoint measurement has displaced derived (or indirect) measurement, I will, second, examine similarities and differences between the two methods. The crucial question I will ask is whether conjoint measurement is an adequate substitute for the measurement of derived quantities. The answer will be in the negative. That is, third, I will argue that the fact that we cannot determine magnitudes of derived quantities directly, i.e., without relying on prior measurement, questions the validity of the assertion that conjoint measurement constitutes a useful alternative to derived measurement.

Measurement as counting of units

According to Narens and Luce (1986) measurement is possible whenever four requirements are satisfied:

> First, the underlying empirical situation is characterized as an ordered relational structure $\chi = \langle X, \succsim, S_1, \ldots, S_n \rangle$, where $\succsim, S_1, \ldots, S_n$ are the *primitives* of the structure (…). These primitives are empirical relations (including possibly operations) on X that characterize the empirical situation under consideration. Second, there are restrictions—axioms—on the structure that reflect truths about the empirical situation. These are to be considered as putative empirical laws. Third, there is specified a numerically based relational structure $\mathcal{R} = \langle R, \geq, R_1, \ldots, R_2 \rangle$, where R is a subset of the real numbers and the R_i, are relations and operations of comparable types to the corresponding empirical ones. Finally, the fourth feature, which accomplishes measurement, is the proof of the existence of a structure preserving mapping from χ into $\mathcal{R}$. We refer to χ as the *empirical* or *qualitative structure*, $\mathcal{R}$ as the *representing structure*, and the structure-preserving mapping as a *homomorphism* or a *representation*. (p. 173)

These four requirements constitute what Narens and Luce (1986) call the "General Representation Theory" (p. 173).

Under scrutiny, however, the general representation theory may require some critical comments: first of all, since numbers do not attach to objects by themselves, what is necessary for the assignment is a so-called *measurement procedure*. Furthermore, "[d]espite the proliferation of measurement axiomatizations" (Krantz et al., 1971, p. 9), the measurement procedure underlying nearly all of them is the so-called "standard-sequence procedure" (p. 6). As Luce and Narens (1994) note: "either directly or indirectly standard sequences are used to establish scales in almost all of the major results of

the FM [i.e., *Foundations of Measurement: Vol. 1*, Krantz et al., 1971]" (p. 225). These comprise all cases of ratio or interval measurement; exceptions being only "purely ordinal cases" (Luce & Narens, 1994, p. 242).

There is a simple explanation as to why the standard sequence procedure plays such a unique role in measurement theory. This is because it is involved both practically and mathematically in constructing the measurement function ϕ or in establishing a homomorphism ϕ as a structure preserving mapping between empirical structures (i.e., magnitudes of quantity) and numerical structures (i.e., real numbers). The procedure is described by Krantz et al. (1971) as follows:

> Select any e in A; this will be the unit. For any other a in A, and for any positive integer n, the Archimedean axiom guarantees that there is an integer m for which $me \succ na$. Let m_n be the least integer for which this is true, namely, $m_n e \succ na \succsim (m_n - 1)e$. Thus, m_n copies of e are approximately equal to n copies of a. As we select n larger and larger, the approximation presumably gets closer and closer and, assuming that the limit exists, it is plausible to define
>
> $$\phi(a) = \lim_{n\to\infty} m_n / n \text{ . (p. 75)}$$

The task of axiomatic measurement theory is to mathematically demonstrate that such a limit exists. The credit to have first solved this task belongs to Otto Hölder (1901). In essence, he proved that the assignment of numbers to objects by means of the standard sequence procedure is legitimate if it satisfies two properties. First, the numbers assigned by counting units in a standard sequence preserve the observational order: i.e., $b \succ c$ if and only if $\phi(b) > \phi(c)$. This is because, "[i]f $b \succ c$, then for some sufficiently fine-grained standards sequence based on some a, we have $b \succ na$ and $na \succ c$, so that $\phi(b) > n\phi(a) > \phi(c)$" (Krantz et al., 1971, p. 4). Second, the numbers assigned are additive with respect to concatenation, so that $\phi(b \circ c) = \phi(b) + \phi(c)$. "The reason is that if n copies of a must be concatenated to approximate b and n' copies to approximate c, then the concatenation of $n + n'$ copies of a will approximate the concatenation of b with c" (p. 4). In simple terms, Hölder demonstrated that numbers and quantities have the same structure.

Furthermore, what is important to keep in mind is that the operation of concatenation need not be interpreted as physical addition (e.g., concatenating rods by laying them end-to-end in a straight line). This applies only to extensive quantities. In the case of difference measurement, for example, what is concatenated are equal end-to-end intervals $a_2a_1 \sim a_3a_2 \sim \ldots$. That is, while "[i]n extensive measurement, a standard sequence has the form a, $a \circ a$, $a \circ a \circ a$, … ; in difference measurement, a standard sequence has the form a_1, a_2, a_2, … where $a_2a_1 \sim a_3a_2 \sim \ldots$. Thus, it is natural to identify a_2a_1 and a_3a_2 each with a, and the overall interval a_3a_1 with $a \circ a$" (Krantz et al., 1971, p. 143). Similarly, in the case of conjoint measurement "the entities that can be concatenated are *intervals within one factor*" (Krantz et al., 1971, p. 18). "[S]cale values are obtained by counting the number of equally spaced levels along each one of the factors" (p. 424).

As pointed out, what is involved mathematically in constructing a numerical assignment ϕ also applies to empirical practice. For instance, Krantz et al. (1971) describe the standard sequence procedure by the example of length measurement as follows:

> A meter stick graded in millimeters provides, in convenient form, the first 1000 members of a standard sequence constructed from a one-millimeter rod. If we observe that rod b falls between

> *na* and (*n+1*)*a*, say, between 480 and 481mm, then we assign it a length between $n\phi(a)$ and $(n + 1)\phi(a)$ (in the present example, between $480\phi(a)$ and $481\phi(a)$, where $\phi(a)$ is the number assigned to a one-millimeter rod and its copies). The value of $\phi(a)$ depends on the selection of a particular rod (say, *e*) to have unit length. If $e \sim ma$, then $\phi(a) = 1/m$. Thus, if *e* is the meter stick, then $m = 1000$ and the length assigned to *b* must be between 0.480 and 0.481 meters; if *e* is a centimeter rod, then $m = 10$ and $\phi(b)$ must be between 48.0 and 48.1 cm. (p. 4)

Since the standard sequence procedure basically consists in counting units it is alternatively called the "counting-of-units procedure" (p. 6).

Note that measurement based on the counting-of-units procedure is identical to the pre-representational, classical understanding of measurement according to which *measurement can be defined as the determination of the ratio of a magnitude of a quantity to another magnitude of the same quantity called unit*. That is, if we find that $na \approx me$, then we have $a/e \approx m/n$. If we take $e = 1$ as unit, we get $a \approx m/n$ (for details on the classical view see Michell, 1990, Chapter 3; Michell, 1999, Chapter 2). That the two definitions are equivalent in meaning follows from the fact that the concept of "quantity" is synonymous with the concept of a "standard-sequence" as a series of units. The conceptual equivalence is acknowledged by Luce and Narens (1994). They note that, "since the process of measurement through standard sequences is usually taken as paradigmatic of 'measurement processes', almost all of the results of FM [i.e., Vol. 1 of *Foundations of Measurement*, Krantz et al., 1971] are valid not only from the RTM [i.e., the Representational Theory of Measurement] viewpoint but from a number of different perspectives about what measurement is" (p. 225). Subsequently, if used without qualification, "measurement" will always mean measurement on ratio or interval scale level. Note that if measurement is conceived as the determination of ratios of magnitudes, then, in contrast to Stevens' (1946) proposal, measurement on nominal and ordinal scales is already about relations between magnitudes of quantity (i.e., equivalence and order relations).

At this point it seems appropriate to explicate the role of axioms of measurement. In analyzing the standard sequence procedure "one is led to the following question: *What basic assumption must be satisfied by* $\succ$ *and* $\circ$ *in order that the standard-sequence procedure can be carried through in a self-consistent manner*" (Krantz et al., 1971, p. 6). The logical step of making these assumptions explicit is called "the *axiomatization* of the measurement procedure" (p. 7). For example, the Archimedean axiom "says that for any *b*, the set of integers *n* for which $b \succ na$ is finite" (p. 25). That is, in the case of length, the Archimedean axiom asserts that there is no smallest length, respectively, that any length *b*, no matter how small, can always be approximated by a standard sequence *na*. Other axioms formulate more "simple" demands the standard sequence procedure must satisfy. For example, the equivalence relation is defined as reflexive ($a \sim a$), symmetric (if $a \sim b$, then $b \sim a$), and transitive (if $a \sim b$ and $b \sim c$, then $a \sim c$). That is, reflexivity implies that in order to construct a standard sequence it doesn't matter which end of a rod is used for the concatenation, symmetry specifies that it doesn't matter in which order the rods are concatenated and transitivity demands that if rod *a* and rod *b* are of the same length and if the same applies to *b* and *c*, then we must find that *a* and *c* are also of equal

length (i.e., what is demanded is the preservation of length). In short, *axioms of measurement are deduced from a logical analysis of the standard sequence procedure*.

According to representational measurement theory, the proposed course of action is to begin with empirically testing axioms before constructing measurement scales (e.g., Krantz, 1971; Krantz et al., 1971, pp. 26–31). Michell (1999) calls the first step the "scientific task" (p. 75). The second step consists of the "instrumental task" (p. 75) of constructing scales or measurement instruments. Note that, as explained, testing axioms of quantity does not lead per se to measurement; what is required in addition is the application of a measurement procedure. Conversely, however, if a standard sequence is successfully constructed, the axioms of quantity are implicitly verified. This should come as no surprise, since, as pointed out, they originate from a logical analysis of the concept of quantity. Hence, if the objective is to obtain measurement on ratio or interval scales, testing axioms prior to the construction process is an unnecessary step. In my view, it is therefore more reasonable to begin directly with the instrumental task. Thus, *the empirical demonstration of measurability and the construction of a measurement device take place simultaneously*. This actually is the standard course of action in physics. The nature of the *co-dependence* between the scientific and the instrumental task will hopefully become evident subsequently, in particular through the example of the establishment of length measurement by means of the three-plate method.

Before concluding, it should be noted that Krantz et al. (1971) describe the so-called procedure of solving inequalities as an alternative for determining ratios between magnitudes; "where it is impractical to go through the elaborate process of constructing standard sequences" (p. 5). Accordingly, Krantz et al. differentiate "between two types of experimental designs: constructive and factorial" (p. 424). While in a constructive design one constructs standard sequences, in a factorial design "a fixed finite set of levels of each factor is selected for the study" (p. 425). In essence, the procedure of solving inequalities consists in setting up inequalities on the basis of empirical observations and in finding a solution to the system. Note that thereby "the concatenation operation $\circ$ is translated into addition + of real numbers, and the observational order $\succ$ is translated into the order $>$ of real numbers" (p. 5). As Krantz et al. point out, the translation uses the two above-described properties characteristic of numerical assignments by the counting-of-units procedure. That is, the translation is legitimate only under the assumption (or hypothesis) that the investigated attributes are decomposable into standard sequences (i.e., a series of equal magnitudes of quantity) or, in short, that they are quantities. In other words, in setting up inequalities one assumes that equal distances between magnitudes of quantity are represented by "equal distances" between numbers. Hence, just as is the case with the standard sequence procedure, the assignment of numbers to objects is eventually accomplished by counting units. However, what is important to understand is that, as will be argued, the problems with conjoint measurement do not actually start with the construction of standard sequences, but with the identification of magnitudes of derived quantities.

Derived versus conjoint measurement

Until the advent of conjoint measurement, the prevailing view in measurement theory was that fundamental measurement necessarily involves operations of physical addition

(Campbell, 1920). In reaction to this view, Luce and Tukey (1964) introduced conjoint measurement as a new method of fundamental measurement which relies at most on operations of ordering between magnitudes of quantity. Within representational measurement theory the development of conjoint measurement had far-reaching consequences. As Berka (1983) notes: "In the initial stages of the evolution of the formal theory of measurement, thought was also given to a derived numerical assignment, yet in the later development of this conception, derived measurement, has been replaced by the theories of the so-called *conjoint* measurement" (p. 125). Thus, while Suppes and Zinnes (1963) still devote a separate section to derived measurement, the topic is almost completely omitted in the *Foundations of Measurement* (Krantz et al., 1971). A comparative analysis of the two methods shall clarify to what extent the replacement is justified.

Before proceeding I would like to make the following conceptual clarifications: basically the term "variable" will be understood as a symbol that represents a quantity. The often used terms "independent variable" and "dependent variable" will be understood as they usually are, i.e., as standing for cause and effect. (However, the notions "latent variable" and "manifest variable," as they are employed in structural equation modeling, will be avoided because of the undesired connotations or the "mythology"; Maraun & Halpin, 2008, p. 114, attached to them. Accordingly, despite the overlap in meaning, the notion "derived quantity," as is defined below, is not considered as equivalent in meaning to the notion "latent variable.") In the case of laws of nature the variables in a function $y = f(x)$ represent magnitudes of physical quantities. If the quantity is, for example, temperature T, then a magnitude of quantity is a particular temperature (e.g., 273.16 as the measurement value for the triple point of water on the Kelvin scale). Quantities will be depicted by capital letters and magnitudes (or levels) of quantity by lower letters. The discovery of laws of nature or of *empirical laws* basically consists of the empirical investigation of how variations in magnitudes in one quantity are *causally* or *structurally* connected to the variations in magnitudes in another quantity, which is a function of the first. An example of a causal law is $PV = T$, where P is pressure, T is absolute temperature, and V is volume. An example of a non-causal or structural law is $D = M/V$, where D is density, M is mass, and V is volume. Although the following argumentation will focus on causal laws, the objections raised will in essence apply to both.

Derived measurement is applicable in situations in which the empirical relation between a quantity P and at least two other quantities A and X can be represented by a non-interactive function f. The most simple of such functions are of the additive form $P = A + X$ and the multiplicative form $P = A \times X$. These are, as Michell (1990) explains, the simplest non-interactive functions from an infinitely large set. He notes: "Unless there are special reasons for not doing so, the noninteractive relationship between P, A, and X may always be expressed in the simple additive or multiplicative forms" (p. 77–78). Sixtl (1982, p. 28) emphasizes that the set of what he prefers to call decomposable functions allows an *algebraic separation* of the variables involved. Obviously, an empirical relation of the form $P = f(A, X)$ can be represented by a non-interactive function f only if algebraic separability is matched by empirical separability. Note that in the case that one of the quantities is maintained at a constant level the relation between the remaining two becomes one of *proportionality*.

Necessary conditions for the applicability of derived measurement are, first, that at least one of the quantities is measurable and, second, that it is possible to determine

magnitudes of quantity for the other two quantities. Let *a, b, c*, … be magnitudes of *A* and *x, y, z*, … be magnitudes of *X*, so that *ax, ay*, … , *cy, cz*, … are magnitudes of *P*. The empirical relation between *A, X*, and *P* can be represented by a non-interactive function *f*, if the effects of *A* on *P* are *empirically separable* from the effects of *X* on *P*. That is, it must be demonstrated that "the two components contribute their effects independently to the attribute in question" (Krantz et al., 1971, p. 247). More precisely, what we must find is that for any combination of two magnitudes *a* and *b* of *A* with magnitudes *x, y, z* … of *X* the values of the ratios $ax/bx = ay/by = az/bz$ … are *constant* or *invariant*. If the separation is empirically possible then the effects of *X* on *P* should cancel out, thus giving the value of the ratio $a/b = c$, where *c* is a constant. (The same reasoning applies in analogy to the measurement of *X*.) In view of this it is evident why *derived measurement can be defined as the discovery of constants in an empirical law* (Ellis, 1966).

What in theory seems simple is in practice usually a highly complex issue. For instance, initially the experiments performed by Boyle (1662) and Gay-Lussac (1802) seemed to demonstrate that the behavior of all gases under heat can be characterized by the relation $PV/T = const$. But attempts at replication in the first half of the 19th century failed, i.e., the more rigorous and technically sophisticated experiments performed by Regnault (1847) and others demonstrated that gases deviate significantly from proportionality (Mach, 1896; Ostwald, 1894). The crucial discovery to advance thermometry over this hurdle was made by Regnault (1847) in the context of his compressibility studies in which he systematically investigated deviations of gases from the laws of Boyle ($PV = const.$) and Gay-Lussac ($V/T = const.$). What he observed was that different gases approach these laws as the pressure is reduced towards the limit of zero. This discovery led to the notion of the perfect or ideal gas and in consequence to the concept of $PV = T$ as "a *limit law* which applies to gases in a state of extreme dilatation; but which deviates all the more from reality the more the gases are compressed, in other words, the closer their molecules move together" (Regnault, 1847, p. 120).[2]

Based on Regnault's discovery, Berthelot (1907) devised a convenient method to measure temperature on a ratio scale, which represents to this day one of the favored methods in gas thermometry (Benedict, 1984; Wensel, 1941). In essence, Berthelot's method relies on the Amagat diagram, i.e., a graphical representation of isothermals of the pressure-volume product *PV* as a function of pressure *P*. The advantage of the graphical representation is that the isothermals of an ideal gas are represented by horizontal lines. Thus, deviations from ideality are easily noticeable and can mathematically be dealt with. Empirically, Berthelot's method requires *P-V* observations at successively lower pressures along two isothermals. Linear extrapolation to zero pressure yields the *PV* intercepts. Thus, in principle, the ratio between any two temperatures can be determined or, more precisely, as Benedict (1984) explains:

> Once a definite number is assigned to one arbitrary state (as 273.15 K to the ice point) or once a definite temperature difference is assigned between two reproducible reference states (as 100 to $T_{steam} - T_{ice}$), all other temperatures on the absolute scale can be determined in principle. In practice, however, no continuous absolute scale is forthcoming, since only a finite number of reliable fixed-point environments exist where such temperature ratios can be defined. (p. 20)

Note that the impediments to the construction of a continuous scale are not theoretical but empirical in nature. They result from limits set by nature to experimental manipulability. What further complicates matters are of course systematic disturbances, i.e., the inadequacies of the equipment used (Childs, Greenwood, & Long, 2000). They may comprise the thermal expansion of the gas bulb, the absorption of impurities in the gas, the difference in the gas at different levels in the pressure sensing tubes, etc. (for details see also Guildner & Thomas, 1982).

Similarly, conjoint measurement is applicable in situations where a quantity P is empirically related by a non-interactive function f to two other quantities A and X. However, the great advantage is seen in the fact that in order to attain measurement none of the quantities must be measurable. What is solely required is that we are able to determine magnitudes of A, X, and P and to apply the operation of ordering to P. Theoretically, the method allows the construction of standard sequences on each factor, so that measurement values can be obtained simply by counting the number of equally spaced levels along each one of the factors.[3]

Following Krantz (1964), the basic idea of the construction of a standard sequence by means of conjoint measurement can be outlined as follows: let us select $ax < bx$ as unit: if some y can be found such that $ay = bx$, then a shift from a to b produces the same change in P as a shift from x to y. Under the condition that the joint effects of A and X on P are additive, the difference between (by - ax) is twice as large as the differences between (bx - ax) and between (ay - ax). If there exists some c and z for which $az = by = cx$ then c and z produce twice the difference from ax than b and y produce from ax, etc. Krantz concludes:

> Thus, by matching changes produced by varying the level of one factor with changes produced by varying the level of the other, and by considering the contributions of the two factors as additive, one obtains a scale on each factor, with scale values summing to give a scale for the quantity being measured. (p. 249)

But, as Krantz et al. (1971), point out, before a standard sequence can be accepted as "equally spaced," the assumption of additivity must be tested empirically, since if "A_1 represents a finite set of levels of some factor and A_2 represents a different factor, there is no reason whatsoever to suppose that when we move from (b_1, b_2) to the next higher level of A_1, say (a_1, b_2), the effect is *exactly* the same as when we move to the next higher level of A_2, say (b_1, a_2)" (p. 20). Hence, the construction process allows an empirical test by entailing a prediction. That is, if we start by selecting level a of A and level x of X and if we select as the next highest level y of X, then we are constrained to select the next highest level b of A so that $ay = bx$. Next, we are forced to select the next level c of A so that $cx = by$ and, similarly, z of X so that $by = az$. "But now, with all degrees of freedom gone," we are forced to have $cy = bz$ "which, empirically, could be false" (p. 21).

It should be noted that there is an intrinsic relation between conjoint measurement and the concept of indifference curves as the central characteristic of derived measurement. Narens and Luce (1986) elaborate:

> The factorizable orderings are very closely related to the concepts of trade-offs and indifference curves that are widely used throughout science: in each case, the equivalence part of the

> ordering describes the trade-off between the factors that maintains at a constant value the amount of the attribute in question, be it mass, loudness, or preference. (p. 169)

Hence, conjoint measurement has much more in common with derived measurement than one might expect. Michell (1999) notes: "a trade-off between equal increases in two attributes identifies equal ratios directly. An increase from, say, X to Y, within an attribute, not only identifies a difference ($Y - X$), but also identifies a ratio (Y/X), the factor by which X is multiplied to reach Y" (p. 204). That is, "[i]dentifying ratios directly via trade-offs results in the identification of multiplicative laws between quantitative attributes. This fact connects the theory of conjoint measurement with what Campbell called derived measurement" (p. 204).

However, though *prima facie* conjoint measurement may seem less demanding than derived measurement, in practice the construction of a standard sequence may be quite difficult to implement, since it places even higher demands on our ability to control phenomena than derived measurement. One of the main causes of concern is the so-called solvability condition, which is a necessary requirement not only in theory but also in the construction process. In conjoint measurement, this axiom demands "that for any ap in $A_1 \times A_2$ and for a certain b in A_1, there exists q in A_2 such that $ap \sim bq$" (Krantz et al., 1971, p. 423). Note that, for example, in the case of temperature measurement—given the problem of establishing fixed-points—the implementation of a constructive design would be practically impossible. That is, it is not possible to arbitrarily select fixed temperature points—as the solvability condition demands—in such a way that in combination with different values of pressure equal intervals of volume expansion can be constructed.

At this point supporters of the representational measurement theory may concede that the constructive design is indeed "difficult to execute because a different construction is required for every subject. Another [reason] is that the construction requires a great deal of care because random error is magnified at each successive stage; moreover, time and order biases may introduce systematic error as well" (Krantz et al., 1971, pp. 424–425). As mentioned earlier, under these circumstances the application of the factorial design is suggested as an alternative measurement procedure. One of the main reasons for the development of measurement inequalities is the elimination of the solvability condition. However, in my view this approach is—relative to derived measurement—no less if not even more demanding than the constructive design. Apart from the theoretical "problem of formulating necessary and sufficient conditions for the existence of a linear representation (i.e., additive or subtractive) for finite data structures" (Krantz et al., 1971, p. 426), there is the practical problem of controlling the phenomena to such an extent that a "set of levels of each factor [i.e., a set of magnitudes of quantity]" (Krantz et al., 1971, p. 425) can be determined free of systematic error. Accordingly, Krantz et al. (1971) note:

> An attempt to apply a finite linear measurement model to empirical data is all too often confronted with the unfortunate situation where the system of equations and inequalities derived from the data (via the measurement model) is inconsistent. In part, at least, this may result from sampling errors that can render the system unsolvable, even if the underlying model is basically valid. (p. 434)

Furthermore, even though "[r]educing error variance is clearly desirable; nonetheless, the model can be so sensitive even to relatively small sampling errors that this approach can turn out to be too costly" (p. 434).

In conclusion, the comparison of the two methods demonstrates that *the attainment of measurement by means of conjoint measurement is more difficult in empirical practice than by means of derived measurement*. In other words, the price paid for making weaker theoretical assumptions (e.g., that none of the quantities must be measurable) are higher demands on our ability to control phenomena. Therefore, whenever at least one attribute is measurable—even though theoretically conjoint measurement would still be applicable—the use of derived measurement is preferable. So far, the identification of magnitudes of quantity (i.e., measurement on a nominal scale) was assumed as non-problematic. In the next section this assumption will be submitted to a critical investigation.

The problem with derived quantities

Krantz at al. (1971) note: "an attribute is called *fundamental* if its measurement does not depend on the measurement of anything else" and they add: "[b]ecause of the inherent logical symmetry of conjoint measurement … all of the traditional physical attributes, are fundamental" (p. 502). The authors explicate this strong statement by referring to "the usual situation in applications of conjoint measurement to physics" (p. 277):

> For example, moving objects can be ordered in three different ways (at least): $\succsim_m$ (mass ordering), $\succsim_v$ (velocity ordering), and $\succsim_p$ (momentum ordering). When we write an object with mass a and velocity q as an order pair aq and consider the momentum ordering $\succsim_p$ on the product set, e.g., $aq \succsim_p a'q'$, we are really *constructing* a product set by observing $\sim_m$ and $\sim_v$ … . Clearly, if there are three orderings, which two we select as independent variables is a matter of convention. (p. 277)

This is also the reason why they agree with Pfanzagl (1968) who expressed the same opinion, namely that no real distinction can be drawn in physics between fundamental and derived measurement. The origin of the faith in the power of conjoint measurement—i.e., the conviction that all kinds of quantity have essentially the same status—rests on the view that magnitudes of any quantity can be deduced "from entirely qualitative observations" (Narens & Luce, 1986, p. 168), i.e., observations which do not require any prior measurement. A critical analysis of the origin of this view will be delivered in the next section. Here, I will focus on the question of whether it is indeed possible to determine magnitudes of any quantity fundamentally.

Luce and Tukey (1964) illustrate the procedure of conjoint measurement by a mechanical example. For instance, they argue, joint effects of mass and gravitational potential difference in producing momentum can be studied with a ballistic pendulum as follows:

> Let a pendulum hanging *in vacuo* be fitted with auxiliary horizontal arms that end in sticky pans, and arrange it so that pairs of spherical pebbles of the same material can be dropped on the pans *simultaneously* from repeatable points of release. We record, qualitatively, the altitude of release and identity of each pebble and the direction of the first swing of the pendulum. Such

> a device is, in essence, a two-pan ballistic pendulum that permits us to compare momentum transfer. If *A* and *B* represent altitudes of release, or, more precisely, differences in gravitational potential between the release points and the pans, and *P* and *Q* represent masses of the pebbles, then the device allows us to compare directly the effect (*A, P*) with the effect (*B, Q*) when the two pebbles are dropped simultaneously. (pp. 4–5)

Of course, Luce and Tukey are aware of the fact that physical quantities can be measured in conventional manners. Accordingly, they "do not claim that conjoint measurement supersedes classic measurement by concatenation, but only that neither is more fundamental than the other" (p. 5).

Admittedly, in the case of length, measurability is entirely deducible from qualitative observations. However, as Dingler (1925) notes, the real problem with empirically testing length for measurability is the problem of circularity. That is, in empiricist theories such as the representational measurement theory, it is usually suggested that the axioms of measurement can be verified by means of straight rods (e.g., Krantz, 1971). But note that straight rods are rigid physical objects of definite geometrical form (i.e., rectangular cuboids or cylinders). Hence, if the empirical testing fails one can always argue that it was due to systematic disturbances, i.e., that the rods were not straight and/or rigid enough. Therefore, one must make sure beforehand that the objects used are suited for the purpose of empirical testing. For instance, how can we ensure that the opposite edges are of equal length? Obviously, if one will want to examine the objects by means of a ruler this will lead automatically into a vicious circle. After all, a ruler is nothing but a straight and rigid rod. Hence, the crucial question is: how can we solve the problem without presupposing that it has already been solved somewhere else?

As Dingler (1933) points out, the task can be accomplished with the help of the so-called "three-plate method," as used in manufacturing precision tools. The procedure can briefly be described as follows: "One takes *three* steel plates *a, b, c* (naturally already smoothed in a rough sense—though this is not necessary in principle) and polishes these on one another in such a manner that in continuous exchange *a* and *b* are ground against each other, then *a* and *c*, and *b* and *c*" (Dingler, 1933, p. 38, as translated in Dingler, 1936/1988, p. 407; for a slightly different variant of the procedure see Goodeve & Shelley, 1877, pp. 11–16; see also Dotson, 2016). Finally, in similar steps of working the metal and operations of geometrical fitting it is possible to manufacture rectangular rods, i.e., physical objects having the geometrical form of rectangular cuboids (for details see Goodeve & Shelley, 1877, pp. 18–20). Thus, rulers can be manufactured without requiring any prior established length measurement. Note that the method of manufacturing a measuring instrument is simultaneously the method of verification of measurability. That is, for instance, if the three plates do not mutually fit upon each other then we can confidently conclude that at least one is not flat. In conclusion, in the case of length measurement the objects "qualitatively observed" are their geometrical forms. Obviously, the method of qualitative observation works with extensive quantities like length; but how about non-extensive quantities? After all, Krantz (1971) explains: "Extensive measurement has had few applications to fundamental measurement of psychological variables, because no concatenation operations are known for variables such as loudness, utility, intelligence, thirst, or anxiety, which satisfy appropriate qualitative laws" (p. 1428).

As usually understood, extensive quantities are dependent on the amount of matter in an object. That is, if an object is divided into any number of parts, the value of an extensive attribute of the object (e.g., volume) is the sum of the values of its parts. In contrast, non-extensive quantities do not depend on the amount of matter in an object. For example, the temperature of the parts of a divided object is the same as the temperature of the whole object. Krantz et al. (1971, p. 502) agree that the distinction between fundamental and derived measurement would make sense if the former would simply mean "extensive" and the latter "non-extensive." This view apparently contradicts Norman Campbell's (1920) belief that fundamental measurement necessarily requires physical additivity. That is, contrary to Krantz et al., Campbell was of the opinion that quantities which do not sustain the operation of physical addition can be measured only derivatively (for details see Michell, 1999). Since the invention of conjoint measurement, however, there is the general opinion that the distinction between extensive vs. non-extensive quantities does not really represent any limitation to what is fundamentally measurable. The shared view among supporters of representation measurement theory is that the applicability of the "method of directly ordering an attribute" (Krantz et al., 1971, p. 502) is sufficient for attaining the fundamental measurement of any attribute.

However, in my view the problem in psychology is not really with the difference between extensive versus non-extensive quantities, but results from the fact that, *if at all, magnitudes of psychological attributes are only identifiable dependent on causally or structurally associated quantitative indicators*. For instance, it is not only impossible to concatenate two instances of motivation, but it is also not possible to place them, so to speak, next to each other and compare them directly with regard to their magnitude, as is possible with two straight rods. (As the three-plate method demonstrates: no other quantity is necessary to empirically verify the measurability of length.) Psychologists are very much aware of the problem. For example, in their guide for how to measure motivation, Touré-Tillery and Fishbach (2014), point out that as "the psychological force that enables action," motivation "cannot be observed or recorded directly" (p. 328). They elaborate, "[r]esearchers measure motivation in terms of observable cognitive (e.g., recall, perception), affective (e.g., subjective experience), behavioral (e.g., performance), and physiological (e.g., brain activation) responses and using self-reports" (p. 328).

In line with the above definition of fundamental attributes an attribute will be called *derived* if its measurement depends on the measurement of something else. Accordingly, *derived quantities can be defined as quantities for which it is not possible to determine magnitudes of quantity directly*. (Note that fundamental versus derived attributes as here defined should not be confused with "base" versus "derived" units as they are defined in the International System of Units. Second, note that the class of derived quantities is larger than the class of intensive quantities. Thus, both force and temperature are derived quantities; but only the latter is also an intensive quantity. And, third, of course associated indicators may themselves be derived quantities.)

Psychological attributes share the *differentia specifica* of derived attributes with physical quantities usually called "forces" (e.g., accelerative forces, electromotive forces, elastic forces, pressure, gravitation, electric fields, etc.). That is, just as is the case with physical forces, magnitudes of "psychological forces" are *in principle* not directly detectable, but only by means of the effects they produce. Actually, with physical forces the

problem was critically noted almost immediately after Newton (1687/1999) introduced the concept into modern physics (Jammer, 1957). Thus, it was already clearly expressed by d'Alembert (1743) through the example of accelerative forces. He noted that:

> Consequently, they cannot manifest themselves to us but only through the effect they produce by accelerating or decelerating the movement of objects, and we cannot distinguish one from another but only by the law and the known magnitude of their effects, i.e., by the law and the quantity of variation they produce in the movement of objects. (as cited in Jammer, 1957, p. 213)

In psychology the problem of derived attributes is usually described with the help of the conceptual dichotomy "observable" vs. "non-observable." This conceptualization may mislead us into believing that psychological attributes are entities which eventually may become directly observable and thus fundamentally measurable. In my view, however, it makes no sense to conceive of psychological attributes as non-observable entities, just as it makes no sense to conceive of physical forces as non-observable physical objects. Note that even though we are able by now to measure electrical current by directly counting single electrons (Bylander, Duty, & Delsing, 2005) this does not make the electromotive force (or rather potential differences) as the *causa movens* in any sense more "visible" and hence directly measurable. Or, for instance, though it may be possible to measure temperature based on counting atoms (Müller et al., 2010; Sanner et al., 2010), temperature (or rather temperature differences), as the cause of the phenomena observed, is still not directly measurable. Nobody searches for methods for visualizing forces as one searches, for instance, for methods to make atoms visible (e.g., at the SuperSTEM; Engineering and Physical Sciences Research Council, 2015).

Accordingly, in the critical discussion it is generally acknowledged that forces cannot be measured fundamentally or directly. That is, there is no debate among physicists that the "non-observability" of forces may be ascribable to a deficit in our methods of detection (Coelho, 2010; Jammer, 1957). Those who question their status don't suggest solving the problem by developing devices for an effect-independent detection. However, what has been tried with much tenacity was to minimize their role by degrading them to theoretical terms or, even more, to eliminate them from physics altogether. For instance, probably most forcefully, Ernst Mach (1868) attempted to purge the concept of force from mechanics, since it is "unfit to enter an empiricist schema, in which existence must be identified with capacity to have or to produce sense impressions" (Bunge, 1966, p. 589). According to Jammer (1957), the "process of eliminating the concept of force from mechanics" (p. 241), as started by Ernst Mach, has been completed in its logical development. In contemporary physics, Jammer claims, force merely "plays the role of a methodological intermediate" (p. 244). Thus, within the framework of classical mechanics, force *denotes* nothing more than a configuration of "gravitational masses, electric charges, magnetic moments, and so forth" (p. 244) and, in consequence, "the product of the inertial mass m of our test body A and its acceleration a, that is, ma, is a function ϕ of the total configuration under discussion" (p. 244). However, what has also become evident in the meantime is that radical attempts to completely purge physics of the concept of force or "non-observable" quantities in general must be considered as failed. As Stepin (2005) recently remarked: "The strict requirement to eliminate non-observable quantities from

theory has never been applied in physics" (p. 273). Above all, the philosophical discussion has had as good as no influence on the practice of measurement (e.g., see the guide for the measurement of force published by the National Physical Laboratory, 1998).

It should be noted that similar eliminative attempts in psychology also completely failed. Most notably is, of course, the radical attempt initiated by Watson (1913) to reduce psychology to "facts of behavior" (p. 159); a research program which came to be known as "methodological behaviorism" (Graham, 2015). One of the main reasons why behaviorism ultimately failed was because it could not deal with "intervening variables," which were first introduced—or rather reintroduced—by Tolman (1938) into psychology as standing for non-observable, psychological factors (e.g., appetite, demand, skill, etc.). Hence, just as in physics, it turned out that empirical phenomena cannot be adequately described without the help of intervening variables or derived quantities.

In conclusion, *there is in principle no observational setup through which psychological attributes will ever become directly detectable and therefore directly measurable*. In this regard behaviorists were certainly justified in their mistrust of "mental entities." In the best case, neural correlates of psychological attributes may someday be identified unequivocally. Thus, similar to the measurement of derived quantities on the atomic level, magnitudes of psychological attributes may theoretically be determined by, say, using as indicator the number of active brain cells in a certain region of the brain. Still, note that, even under these ideal circumstances, the *proportionality* between magnitudes of motivation M and number of active brain cells N would have to be first established experimentally following the procedure of derived nominal measurement described below. Merely observing brain activity will not tell us which psychological phenomenon it is associated with, just as observing moving elementary particles will not tell us if the cause of movement is electric, thermic, or mechanical in nature. But doesn't this place psychological attributes "in a sort of never-never land—a domain which is forever inaccessible to scientific inquiry" (Krech, 1950, p. 284)? Before answering this question the following issue must be clarified: given the substantial problems described, why is the detection of magnitudes of derived attributes not perceived as problematic? This topic will be dealt with in the next section. The subsequent section will be devoted to the question of how—despite the fact that they are not directly detectable—magnitudes of derived quantities can still be identified.

Humans as measurement instruments

In my view the reason for the misapprehension of the problem of derived attributes can be traced back to the widespread view that test subjects are somehow capable of determining magnitudes of quantity of psychological attributes. That is, psychological tests usually employ some form of a psychometric scale (e.g., the Likert scale) to collect data. Note that if the data are considered to satisfy the demands of interval or ratio scales, the hand of the test subject moving a pencil along the scale and making a mark is treated as equivalent to the movement of a pointer in the display of a measuring device. Hence, there is legitimacy in saying that the unit "test subject-rating scale" is quite literally regarded as an embodiment of a measurement instrument.

More precisely, a thus conceived test–testee unit can be viewed, as was already noted by Suppes and Zinnes (1963), as an instance of "pointer measurement," i.e., as "a numerical assignment (either fundamental or derived) based on the direct readings of some validated instrument" (p. 20). Note that by holding this view one is implicitly obliged to assume "that someone has taken the trouble to verify that the deflections of the pointer under certain 'standard' conditions do indeed correspond to the values of a given fundamental or derived numerical assignment" (p. 20). However, instead of "pointer measurement," the term more commonly used in psychology is "*measurement by means of convention (by fiat)*" (Berka, 1983, p. 131); a concept that has been introduced into the social sciences by Torgerson (1958). In essence, measurement by fiat is "based on the *belief* that the respective attribute is measurable, and that tests lead to measurement on interval scales" (Orth, 1974, p. 41). Berka (1983) critically notes: "Since the measurement on the basis of conventions depends on the 'intuition of each individual experimenter', the results obtained via this pseudo-measurement are considerably controversial" (pp. 131–132). Suppes and Zinnes (1963) likewise observe: "All too often in the behavioral sciences a direct reading instrument is available (and used) despite the fact that its readings are not justified; the readings do not correspond to any *known* fundamental or derived numerical assignment" (p. 21). Despite this deficiency, the practice continues unabated to the present day. As Morris, Grice, and Cox (2016) recently pointed out: "It is common practice in psychology to devise 'measurement' procedures by imposing rating scales (e.g., Likert items) onto phenomena and treating the values they produce as quantities" (p. 1).

Importantly, however, even where the quantitative hypothesis is empirically explored, the view that humans have the capabilities of measurement instruments enters the investigation as a background assumption or an auxiliary hypothesis. For example, consider Michell's (1994) investigation of the hypothesis "that differences between distinct attitudes on the same dimension are quantitative" (p. 244) on the basis of Coombs' theory of unfolding (Coombs, 1964; see also Michell, 1990, Chapter 7). In order to apply the method of unfolding one has to identify points (or magnitudes of quantity) on the dimension investigated. These are of course not determined directly, as is in general the case with derived attributes, but indirectly by asking individuals i if they prefer x to y. Also, note that, though on the face of it only judgments about ordinal relations are demanded of the test participants, in reality it is implicitly assumed that the responses are determined by consulting an "internal" scale on which the distance of x and y to the so-called point of maximum preference is estimated. In addition, it is hypothesized that all participants agree about the location of the items on the quantitative dimension; in other words, they are not only conceptualized as valid, but as calibrated instruments as well. Hence, as is usually the case in psychology, "test–testee units" are conceived as instances of pointer measurement.

In conclusion, the fact that the problem of derived quantities does not attract the necessary attention in psychology can be attributed to the commonly shared, unquestioned assumption that humans are capable of unequivocally identifying magnitudes of quantity, i.e., that they are capable of performing the tasks usually ascribed to measurement instruments. I believe that it is this view of "humans as measurement instruments" that creates in psychology the illusion that magnitudes of derived quantities can be determined fundamentally. However, by relying on the assumption that in test subjects "a

direct reading instrument is available" (Suppes & Zinnes, 1963, p. 21) the problem of the determination of magnitudes of derived quantities—the task which actually must be first and foremost dealt with—is evaded. In other words, it is hypothesized that the problem has already been solved "within" the test participants and that consequently we only have to ask them about the location of the items on the quantitative continuum. The nature of the illusion becomes, I believe, even more evident if one considers that the "method of asking" the objects under investigation for the location of magnitudes on a quantitative continuum is inapplicable in physics, since physical objects are not able to reply to our questions or follow our instructions. (Note that the logical possibility that humans may have the capabilities of measurement instruments is not disputed; though it is in my view—for reasons stated elsewhere; Trendler, 2009—an unrealistic hypothesis.)

As the effect, so the cause

In experimental practice, the problem posed by derived quantities is not so much a conceptual problem, but rather a problem of empirical manipulability. Accordingly, I believe that *the solution to the problem of derived quantities is not theoretical, but practical in nature*. In other words, the problem of derived quantities—as a "problem of theoretical terms" (Holger, 2017)—has a practical solution. The following brief analysis of the measurement of "forces" in physics should help to explicate how the problem is circumvented in practice. This is all the more relevant since psychological factors are usually conceived—similarly to forces in physics—as the "force" or the "drive" behind observed behavior.

The category of what Fuchs (2010) calls "driving forces" is probably one of the largest subsets of derived physical quantities. To start with, let us consider, for instance, the phenomenon of pressure equalization. It is experimentally demonstrable with the help of two communicating tanks connected by a lockable hose at their bottoms and filled with oil. If the level of fluid in the tanks is different, we will observe that, as the level in one of the tanks decreases it will increase in the other tank until the system reaches equilibrium. Fuchs explains: "There is *dynamics* as long as we have difference of levels in the two tanks—the level difference is conceptualized as the *driving force* of the flow of fluid" (p. 18). The driving force in a hydraulic system is called pressure. In analogy, in electrical circuits the electric charge flows from higher to lower electrical intensity and the driving force is called electric potential or electromotive force. "Ohm's law actually establishes the relationship between the flux of charge, the properties of the conductor, and the electrical driving force responsible for the flow" (p. 84). Furthermore:

> We say that heat flows from the hotter body of water to the colder one as long as there is a temperature difference. We interpret temperature differences as the *driving force for the flow of heat and temperature as the thermal level* because the behavior of the temperatures resembles that of water levels in communicating tanks, or of voltages of capacitors connected by a resistor. (p. 100)

However, the crucial point is that directly detectable phenomena alone "do not tell us anything about why water should be flowing at all. In electrical circuits as well, we need a quantity which is responsible for setting up currents of charge in the first place" (p. 28).

Hence, *derived quantities are necessary for a complete description and explanation of the phenomena observed*.

But how can we manipulate and control driving forces if they are not directly identifiable? As argued elsewhere (Trendler, 2009), the discovery of causal or structural relations is a matter of apparatus construction. In the case of temperature, we need a heat engine which generates the thermal energy necessary to bring the system to a higher or lower temperature. In a hydraulic circuit the task of manipulation and control of pressure can be accomplished by means of a pump; in the case of an electrical circuit voltage can be manipulated with the help of a battery. The construction process is normatively guided by the idea of *proportionality*. Robert Hooke (1678) pointedly expressed the idea with reference to the measurement of force by means of the law of the spring: "*Ut tensio sic vis*; That is, The Power of any Spring is in the same proportion with the Tension thereof: That is, if one power stretch or bend it one space, two will bend it two, and three will bend it three, and so forward" (p. 1). Whewell (1840) generalized the idea into one of his "[a]xioms which relate to the idea of cause" (p. 169), namely: "*Effects are proportional to their Causes, and Causes are measured by their Effects*" (p. 171).

Of course, causal or structural proportionality cannot be assumed a priori as empirically true, but must be demonstrated experimentally. It depends on the ingenuity and persistence of the experimenter to establish processes which realize proportionality. The method essentially consists in using the quantitative indicator P to identify magnitudes of the derived quantity A on the basis of Mill's method of concomitant variations, which states: "*Whatever phenomenon varies in any manner whenever another phenomenon varies in some particular manner, is either a cause or an effect of that phenomenon, or is connected with it through some fact of causation*" (Mill, 1843, p. 470). The criterion for the successful identification of magnitudes of derived quantities (e.g., temperature) by means of already measurable quantities (e.g., volume) has been clearly specified by Regnault (1847, pp. 164–165). That is, for example, in the case of gas thermometry it must be empirically demonstrated that under identical circumstances T (e.g., the triple point of water) the investigated gas always occupies the same volume V and, moreover, that this outcome is arbitrarily *reproducible* with the same and with different experimental apparatus. In other words, *ceteris paribus*, we must always obtain, in the limits of random error, identical *measurement values* for V. Thus, the known value of the magnitude of the quantity V is taken to indicate a particular though yet unknown value of magnitude of the quantity T. In the nomenclature of the representation measurement theory this might be called *derived nominal measurement*. In brief, "the effect is an unfailing index of the amount of the cause; and if it be a measurable effect, gives a measure of the cause. We can have no other measure; but we need no other, for this is exact, sufficient, and complete" (Whewell, 1840, p. 172).

At this point it should be mentioned that, in my view, the method of derived measurement represents a solution to the problem identified by Chang (2004, p. 57) as the problem of nomic measurement, and which can be resumed as follows: if X is a not directly observable quantity related to an observable quantity P by a function $f(X)$, then it must be the case that in order to measure X we must know f. But if f is unknown and cannot be discovered empirically—because that would involve knowing the value of both X and P—how is it possible to measure X? Indeed, if only two quantities are taken into

consideration, then X can be measured only on an ordinal scale. What would result at best is what Ellis (1966) calls "associative measurement" (p. 90) or what alternatively may be called *derived ordinal measurement*. With at least three quantities, however, the problem can be solved by means of derived measurement. That is, the solution to the problem of nomic measurement does not result, as Sherry (2011) believes, from the "explanatory and predictive power of concepts that presuppose temperature's quantitative status" (p. 517), but from the discovery of quantitative invariances. That is, *the detection of constants in an empirical law constitutes necessary and sufficient evidence that a derived quantity is not merely measurable on an ordinal, but on a ratio scale*.

Finally, note that, as already pointed out (Trendler, 2013), indicators must not be *continuous quantities* (e.g., length), but can also be *discrete quantities* (e.g., number of electrons). Derived measurement can also be attained by means of probabilities (i.e., distributions of discrete or continuous quantities) as is described in statistical physics (Huang, 2001). This was until recently only a theoretical option, but in the meantime—as a consequence of the phenomenal progress in experimental physics—real applications have become a reality. For example, as mentioned, attempts are now made to measure temperature based on fluctuations in number of atoms (Sanner et al., 2010). The method relies on counting the number of atoms in small probe volumes of cold gas clouds. As Sanner et al. explain: "[m]any iterations allow us to determine the average atom number N in the probe volume and its variance $(\Delta N)^2$. For independent particles, one expects Poisson statistics, i.e. $(\Delta N)^2/\langle N\rangle = 1$" (2010, para. 5). A similar role can be ascribed in psychology to probabilistic models of measurement (Rasch, 1960). In Rasch models, the odds ratio $\theta = p/(1\text{-}p)$, where p is the probability of a correct response, is expressed as a function of A, person ability, and D, item difficulty, i.e., $\theta = A/D$ (Andrich, 1988). As is evident from Rettler's (1993) presentation of Rasch's concept of specific objectivity, measurement on the basis of Rasch models is equivalent to derived measurement. That is, what specific objectivity demands is that, for any two items i the ratio $\theta(i_1)/\theta(i_2)$ is constant independently of which persons k are used to determine the ratio, and vice versa, that for any two persons k the ratio $\theta(k_1)/\theta(k_2)$ is constant independently of the items i used for measurement (see also Rasch, 1977). Hence, Rasch models for measurement are instances of derived measurement.

To conclude, *first, given that magnitudes of derived quantities cannot be determined without the help of quantitative indicators and, second, given the superiority of derived measurement in experimental practice, we must recognize that conjoint measurement is superfluous as a method to demonstrate measurability*. Hence, even if—contrary to the Millean quantity objection—the application of conjoint measurement were possible in psychology, it is not required, since, given the "non-observability" of psychological attributes, derived measurement would be the method of choice for solving the task of measurement.

Escape route: Derived measurement?

Given that conjoint measurement constitutes a poor substitute for derived measurement, one may ask: doesn't derived measurement represent an alternative to conjoint measurement; an escape route, so to speak, for the attainment of measurement in psychology?

After all, it is not the case that psychology completely lacks quantitative indicators. For example, given the variable reaction time R, empirical laws of the form $R = D/A$, where A is person ability and D is item difficulty, could easily be submitted to an empirical verification. Of course, as mentioned, the independent variable might also be a neuronal correlate of ability (e.g., the number N of active neurons). Or, instead of these rather fictitious examples, take Rasch models of measurement by means of which psychologists are currently attempting to establish measurement in psychology.

However, to my knowledge there is no evidence that any constants have been found; neither by testing probabilistic models nor by any other means. Instead, what is usually offered as evidence for the existence of quantitative relations is the statement that a range of data statistically fit the measurement model (e.g., Bond & Fox, 2015; Stenner, Fisher, Stone, & Burdick, 2013). But, as Meehl (1990) notes, statistical significance doesn't tell us anything about reproducibility. He elaborates:

> A scientific study amounts essentially to a "recipe," telling other cooks how to prepare the same kind of cake the recipe writer did. If other competent cooks can't bake the same kind of cake following the recipe, then there is something wrong with the recipe as described by the first cook. If they can, then, the recipe is all right, and has probative value for the theory. It is hard to avoid the thrust of the claim: *If I describe my study so that you can replicate my results, and enough of you do so, it doesn't matter whether any of us did a significance test; whereas if I describe my study in such a way that the rest of you cannot duplicate my results, others will not believe me, or use my findings to corroborate or refute a theory, even if I did reach statistical significance*. So if my work is replicable, the significance test is unnecessary; if my work is not replicable, the significance test is useless. (p. 138)

Hence, by relying on model fit only it is premature to claim that "the quantitative hypothesis is sustained" (Stenner et al., 2013, p. 1). What one should do instead is, as explained above, demonstrate that ratios between particular magnitudes of quantity are invariant. This obviously presupposes the reproducibility of measurement values. Without the empirical demonstration of invariance, the "measures" for person ability A or item difficulty D, constructed from "raw scores" (i.e., for person ability the count of items on which the person succeeds and for test difficulty the number of items which the person fails to pass) by means of a measurement model (Wright, 1997), cannot really be considered "measurement values." Actually, statistical significance does not even guarantee that the calculated "measures" satisfy the properties of an ordinal scale. In general, *we are licensed to speak of measurement proper only if the property of invariance is firmly established as an empirical fact* (for supplementary criteria of measurement see Trendler, 2013).

Given the importance of the topic a brief digression may be permitted. As Wright (1997) notes, ignoring the question of measurability is the reason "why so much social science has turned out to be no more than transient description of never-to-be-reencountered situations, easy to contradict with almost any replication" (p. 35). Hence, it should come as no surprise that psychology is characterized—as has been found by the Open Science Collaboration (2015)—by a lack of reproducibility.[4] The Open Science Collaboration presents the result as normal, as "the reality of doing science" (p. aac4716-7) or as reflecting "a cumulative process of uncertainty reduction" (p. aac4716-7). In my view, however, this is only half the truth; the whole truth is that an empirical science is

not only about an interminable uncertainty reduction—i.e., a never-ending search without ever delivering any concrete or only tentative results—but eventually has to rest on the certain solid bedrock of *procedural knowledge* or *know-how* as the given ground. I believe it is appropriate to say that this basis has something "absolute" about it. That is, contrary to Popper's (1935/2002) view that "[s]cience does not rest upon solid bedrock" (p. 94), that the "bold structure of its theories rises, as it were, above a swamp" (p. 94), *empirical theories are ultimately built over the solid bedrock of replicable empirical findings as the natural base*. A new theory may be necessary, but once established empirical know-how remains unaffected. In this sense, for instance, the three-plate method is absolute; since—though the theory about space has changed from Euclidean to Non-Euclidean—the procedural knowledge or the "recipe" of how to manufacture "true planes" (Whitworth, 1858, p. 3), is irrevocable and indispensable. The bedrock of the physical sciences is quite literally solid rock; since all measurement depends directly or indirectly on length measurement and since reference surfaces "are generally granite instead of cast iron" (Dotson, 2016, p. 303). Solid bedrock is what any empirical science should ultimately strive for.

If it comes to measurement in psychology no such solid bedrock may ever be reached. As argued in detail elsewhere, the testing of the quantitative hypothesis only makes sense under circumstances where it is possible to control systematic disturbances (Trendler, 2009). The problem is invariably solved through the construction of experimental apparatus. This method is inapplicable in psychology, since—contrary to physical processes—it is not possible to capture psychological processes in experimental apparatus, devices, or machines as would be required in order to control systematic disturbances. (Obviously probabilistic models represent no exception to this objection since they only account for random error.) Therefore, I believe that derived measurement does not represent a realistic alternative, neither in the social sciences in general nor within psychology in particular. Actually, the total failure of Hullian behaviorism, which in essence represents an application of derived measurement to animal psychology, already stands witness to the inapplicability of derived measurement in psychology (Koch, 1954).

Conclusion

Historically, the development of conjoint measurement theory can be interpreted as a reaction to Campbell's "dictum that fundamental measurement rests on associative, monotonic operations of combination" (Narens & Luce, 1986, p. 168). As Narens and Luce point out, this conclusion was the result of "a curious debate [that] ensued during the 1920s and 30s about what … is measurable" (p. 168). "The debate reached its intellectual nadir with the 1940 Final Report of a Commission of the British Association for Advancement of Science (Ferguson et al., 1940) in which a majority declared fundamental measurement in psychology to be impossible because no such empirical operations could be found" (p. 168; for details see Michell, 1999).

However, what is rather curious in retrospect is that, although conjoint measurement did not produce any substantial results in half a century, its meaningfulness has yet not seriously been questioned. After all, Luce and Tukey (1964) confidently claimed that the new type of fundamental measurement leads quite naturally "to scales of the highest

repute: interval and ratio scales" (p. 4); but up to this point no such scales have materialized in psychology or anywhere else in social sciences. The claim now resembles an empty promise. What advocates of conjoint measurements fail to notice is that it is not possible to deduce values for any quantity "from entirely qualitative observations" (Narens & Luce, 1986, p. 168). The misapprehension is, as has been argued, the result of misconceiving humans as measurement instruments; i.e., it results from the erroneous view that humans are able to report positions of magnitudes of quantity as if they are reading them off from an "internal" scale. Once this is realized, it becomes clear why the conviction expressed by Michell (1999) that Luce and Tukey "showed that the measurement of derived attributes did not depend upon the prior measurement of any other attributes" (p. 205) has to be regarded as inaccurate. Quite the contrary, as has been argued, *already the determination of magnitudes of derived attributes depends on the measurement of other attributes*.

Suppes and Zinnes (1963) emphasize that it may

> be a difficult mathematical problem to show that a given scale is (or is not) an interval scale, but this is not to suggest that the existence of an interval scale is a matter for philosophical speculation or that it depends on the whims and fancies or even the position of the experimenter. (p. 3)

This is certainly true, but just as true is that the temptation of mathematical "elegance and esthetics" (Krantz et al., 1971, p. 26) as inherent in the representation measurement theory can result in becoming lost in a realm of abstractness; where without the resistance of reality, anything seems possible. Indeed, questions of measurability should not depend on the whims and fancies of the experimenter, but likewise they should not depend on the abstract aloofness of the mathematician. As Kant (1781/1998) famously remarked: "Thoughts without content are empty, intuitions without concepts are blind" (pp. 193–194). Similarly, without the mathematics of measurement the empirical method is blind, but it is just as true that theory without experience is empty. In the end, I believe, we must accept that the symmetry between dependent and independent variables observed in conjoint measurement—and in consequence the view that all measurement is fundamental—is nothing but the result of a mathematical illusion, of the shadows cast by mathematics on reality.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. What has been recognized, according to Luce (2011), is that "we also have the interplay between value and risk, which means that multiplication as well as addition is in play" (Section 2, para. 4). In consequence, this "slight mathematical oversight … has put us on a misguided course for over a century. And that course may, in a number of ways, have been

scientifically misleading" (Section 4). Furthermore, Luce notes: "Individual differences are important. This has significant implications—e.g., many behavioral economic experiments should be redone taking into account individual types" (Section 3.1, last para.).

2. Recently Bringmann and Eronen (2016), following Chang (2004), described Regnault's approach as "anti-theoretical to the extreme" (p. 30). In my view this depiction misrepresents the purpose of Regnault's work. As pointed out by Ostwald (1894), Regnault's investigations on the expansion of gases by heat can be traced back to Rudberg's failure to empirically verify the coefficient of expansion of gases as calculated by Gay-Lussac. Regnault correctly suspected that the error was due to deficiencies in the experimental apparatus used. The circumstances described by Ostwald are certainly one of the main reasons why Regnault put so much effort into apparatus construction and, in particular, why he "studied the whole subject of thermometry critically with the constant care of identifying all sources of error" (Poncet & Dahlberg, 2011, p. 391). To construe Regnault's scrupulosity as a master experimentalist to his detriment is, in my view, unfair at least.

   Second, Regnault did not view the collection of data as an end in itself, as an anti-theoretical approach would imply, but the "aim of performing accurate measurements was to produce useful data as a basis for empirical generalization, thus attempting to solve theoretical problems" (Poncet & Dahlberg, 2011, p. 390). For a proper assessment, it should also be taken into consideration that Regnault's empirical "results were used constantly for most of the century to validate scientific theories and in particular in thermodynamics" (p. 388). Regnault was very much aware that not all problems can be solved completely atheoretically. As Rowlinson (2010) notes, with regard to the deviations of real gases from the ideal gas law Regnault, for instance, "lamented that the matter was too difficult to be solved in the laboratory and urged the mathematicians to give it their attention" (p. 46). Actually, he already hypothesized that the behavior of real gases depends on the behavior of the molecules under heat (e.g., Regnault, 1847, p. 120); an idea taken up by other physicists and developed into the kinetic theory of gases. Hence, Regnault's position with regard to theory is, I believe, more accurately described as frugal. One could say that with regard to theory formation he was guided by a version of Occam's razor or the law of parsimony (Baker, 2016). Psychology in particular seems prone to the danger of ignoring this guiding principle and thus having to bear the consequences. For instance, consider cognitive psychology, where psychologists attempt to unravel complex "internal" systems (e.g., the structure of memory) based on a small "external" basis (e.g., reaction time measurements). As I have argued elsewhere (Trendler, 2013), this endeavor is critically *underdetermined*, i.e., the empirical data are not able to put any meaningful restraint on theory. Cognitive psychologists are not even able to uniquely decide in which format (i.e., digital or analog) information is encoded in the brain. How does one expect any progress if even this most basic requirement is open to discussion? Or to take the example of short-term memory research invoked by Bringmann and Eronen (2016) to support their case; given that unitary-store models may just as well describe the phenomena, there is not even agreement on whether short-term memory really exists (Jonides et al., 2008; on the elusiveness of memory research see also Roediger, 2008; Tulving, 2007).

   Third, given Regnault's predilection for accurate measurements, it is safe to assume that he would not have agreed with Bringmann and Eronen (2016) that "it is possible to do good science based on relatively bad measurements" (p. 32; see also Sherry, 2011). In this regard, it is also misleading to claim that "there are surprising parallels between temperature measurement in the first half of the 19th century and the current situation in psychological research practice" (Bringmann & Eronen, 2016, p. 29). The situation is not even similar to the situation in physics in the 16th century, at the beginning of the development of thermometry, when the first spirit-in-glass thermometer, the so-called little Florentine thermometer,

was manufactured. As has been empirically demonstrated based on a set of surviving instruments, these thermometers already delivered very good, i.e., comparable data, so that it is "possible to interpret readings made with the Florentine thermometers in terms of modern temperature scales with considerable confidence" (Vittori & Mestitz, 1981, p. 118; see also Camuffo, 2002). Nothing similar is to be found in psychology. Actually, *psychologists have not even bad measurement data; they have no measurement data at all.* Hence, the current situation in psychology is rather similar to the situation in physics prior to the pneumatic experiments performed by Philo of Byzantium, at about the end of the second century B.C. (Middleton, 1966), that is, prior to any systematic investigation into and therefore prior to any accurate knowledge about the quantitative nature of any attribute. Therefore, I believe that the situation is incomparably worse in psychology than the situation in which Joseph Black found himself when developing the theory of latent heat relying on "(relatively) bad measurements" (Bringmann & Eronen, 2016, p. 27; see also Sherry, 2011). In consequence, if measurement is really the aim, "[w]hat psychologists can learn from the history of physics" (Bringmann & Eronen, 2016, p. 27) is that they should not proceed with business as usual and thus preserve the *status quo* of quantitative psychology as a pathological science (Barrett, 2018; Michell, 2000, 2008), but they should rather follow Regnault's lead and strive to obtain as good data as possible with as good apparatus as can be manufactured. Otherwise it is advisable to consider procedures of data analysis which are suited for non-quantitative data (e.g., Grice, Barrett, Schlimgen, & Abramson, 2012).

Fourth, Bringmann and Eronen (2016) argue that the "*simple* [emphasis added] gas laws" (p. 31), apparently due to their simplicity, are insufficient for measuring temperature on a ratio scale, since the fixed points of the thermometric gas scale are "conventions based on practical considerations" (p. 31) and, since there is "no plausible theoretical definition for what it means for temperature to change by one degree" (p. 31). That this is incorrect—respectively, that no *complex* theory like thermodynamics is required for the determination of ratios of magnitudes between different temperatures—has hopefully become evident from the presentation of temperature measurement by means of Berthelot's method (for details on different definitions of temperature—i.e., based on the gas laws, on thermodynamics or the kinetic theory of gases—and the equivalence between them see Berthelot, 1907, pp. 4–6).

3. Michell (1990) notes, conjoint measurement may not only "be extended forwards beyond the three variable case (i.e., $P = f(A, X)$) to four or more variable cases (i.e., $P = f(A_1, \dots, A_n)$)" (p. 84), but it may also "be extended backwards to cases involving one variable" (p. 84). Thus, extensive measurement can be conceived as a special case of conjoint measurement. If Q is an extensively measurable variable, then we have $A = Q$, $X = Q$, and $P = Q + Q$. "That is, the values of A and X are the values of the variable Q itself and P are these values as combined under the physical operation of addition" (pp. 84–85).
4. The Open Science Collaboration study takes center stage in the so-called "replication crisis" (e.g., Maxwell, Lau, & Howard, 2015). What is most striking is that, while the debate about the reproducibility of psychological science drew much attention within the psychological community (and even got the attention of the media, e.g., van Bavel, 2016), the measurement debate (e.g., Bringmann & Eronen, 2016)—which takes place quasi simultaneously—does not attract any attention beyond the small circle of scholars interested in the topic of measurement; even though the question of measurability is, in my view, not only the more fundamental topic, but also the main explanation for the failure of reproducibility. It is also noteworthy that the debate about measurability in psychology is not even noticed by authors (e.g., Loken & Gelman, 2017) who are aware that measurement, or more precisely, measurement error, may be the root problem of the failure to replicate.

References

Andrich, D. (1988). *Rasch models for measurement*. Newbury Park, CA: Sage.

Baker, A. (2016, December 20). Simplicity. In E. N. Zalta (Ed.), *Stanford encyclopedia of philosophy*. Retrieved from https://plato.stanford.edu/entries/simplicity/

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

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

Benedict, R. P. (1984). *Fundamentals of temperature, pressure and flow measurements* (3rd ed.). New York, NY: Wiley.

Berka, K. (1983). *Measurement: Its concepts, theories and problems*. Dordrecht, the Netherlands: Reidel.

Berthelot, D. (1907). Sur les thermomètres a gaz et sur la réduction de leurs indications a l'échelle absolue des températures [On gas thermometers and on the reduction of their indications to the absolute scale of temperature]. *Travaux et Mémoires du Bureau International des Poids et Mesures*, *13*, 1–113.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human science* (3rd ed.). London, UK: Routledge. doi: 10.1111/j.1745-3984.2003.tb01103.x

Boyle, R. (1662). *A defence of the doctrine touching the spring and weight of the air*. London, UK: Printed by F. G. for Thomas Robinson.

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

Bunge, M. (1966). Mach's critique of Newtonian mechanics. *American Journal of Physics*, *34*, 585–596. doi: 10.1119/1.1973119

Bylander, J., Duty, T., & Delsing, P. (2005). Current measurement by real-time counting of single electrons. *Nature*, *434*, 361–364. doi: 10.1038/nature03375

Campbell, N. (1920). *Physics: The elements*. Cambridge, UK: Cambridge University Press.

Camuffo, D. (2002). Calibration and instrumental errors in early measurements of air temperature. *Climatic Change*, *53*, 297–329. doi: 10.1023/a:1014914707832

Chang, H. (2004). *Inventing temperature: Measurement and scientific progress*. Oxford, UK: Oxford University Press. doi: 10.1093/0195171276.001.0001

Childs, P. R. N., Greenwood, J. R., & Long, C. A. (2000). Review of temperature measurement. *Review of Science Instruments*, *71*, 2959–2978. doi: 10.1063/1.1305516

Cliff, N. (1992). Abstract measurement theory and the revolution that never happened. *Psychological Science*, *3*, 186–190. doi: 10.1111/j.1467–9280.1992.tb00024.x

Coelho, R. L. (2010). On the concept of force: How understanding its history can improve physics teaching. *Science & Education*, *19*, 91–113. doi: 10.1007/s11191–008–9183–1

Cohen, J., & Cohen, P. (1975). *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Coombs, C. H. (1964). *A theory of data*. New York, NY: Wiley.

d'Alembert, J.-B. R. (1743). *Traité de Dynamique* [Treatise on dynamics]. Paris, France: David l'Aîné.

Dingler, H. (1925). Über den Zirkel in der empirischen Begründung der Geometrie [On the circle in the empirical foundation of geometry]. *Kant-Studien*, *30*, 310–330.

Dingler, H. (1933). *Die Grundlagen der Geometrie* [The foundations of geometry]. Stuttgart, Germany: Enke.

Dingler, H. (1988). Method instead of epistemology and philosophy of science. *Science in Context*, *2*, 369–408. (Original work published 1936)

Dotson, C. L. (2016). *Fundamentals in dimensional metrology* (6th ed.). Boston, MA: Cengage Learning.

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

Engineering and Physical Sciences Research Council. (2015, February 20). SuperSTEM microscope sees single atoms. *Science Daily*. Retrieved from www.sciencedaily.com/releases/2015/02/150220083727.htm

Ferguson, A. C. S., Myers, R. J., Bartlett, H., Banister, F. C., Bartlett, W., Brown, N. R. … Tucker, W. S. (1940). Quantitative estimates of sensory events: Final report of the committee appointed to consider and report upon the possibility of quantitative estimates of sensory events. *The Advancement of Science. The Report of the British Association for the Advancement of Science*, *1*, 331–349.

Fuchs, H.-U. (2010). *The dynamics of heat* (2nd ed.). New York, NY: Springer.

Gay-Lussac, J. L. (1802). Recherches sur la dilatation des gaz et des vapeurs [On the expansion of gases by heat]. *Annales de Chimie*, *43*, 137–175.

Goodeve, T. M., & Shelley, C. P. B. (1877). *The Whitworth measuring machine*. London, UK: Longmans, Green & Co.

Graham, G. (2015, March 11). Behaviorism. In E. N. Zalta (Ed.), *Stanford encyclopedia of philosophy*. Retrieved from https://plato.stanford.edu/entries/behaviorism

Grice, J. W., Barrett, P. T., Schlimgen, L. A., & Abramson, C. I. (2012). Toward a brighter future for psychology as an observation oriented science. *Behavioral Sciences*, *2*, 1–22. doi: 10.3390/bs2010001

Guildner, L. A., & Thomas, W. (1982). The measurement of thermodynamic temperature. In J. F. Schooley (Ed.), *Temperature: Its measurement and control in science and industry* (Vol. 5, pp. 9–19). New York, NY: Reinhold.

Hölder, O. (1901). Die Axiome der Quantität und die Lehre vom Mass [The axioms of quantity and the theory of measurement]. *Berichte über die Verhandlungen der Königlich Sächsischen Gesellschaft der Wissenschaften zu Leipzig, Mathematisch-Physische Classe*, *53*, 1–64.

Holger, A. (2017, July 20). Theoretical terms in science. In E. N. Zalta (Ed.), *Stanford encyclopedia of philosophy*. Retrieved from https://plato.stanford.edu/entries/theoretical-terms-science/

Hooke, R. (1678). *Lectures de potentia restitutiva, or of spring explaining the power of springing bodies*. London, UK: John Martyn.

Huang, K. (2001). *Introduction to statistical physics*. London, UK: Taylor & Francis.

Jammer, M. (1957). *Concepts of force: A study in the foundations of dynamics*. Cambridge, MA: Harvard University Press.

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*, 193–224. doi: 10.1146/annurev.psych.59.103006.093615

Kant, I. (1998). *Critique of pure reason*. Cambridge, UK: Cambridge University Press. (Original work published 1781)

Koch, S. (1954). Clark L. Hull. In W. K. Estes (Ed.), *Modern learning theory: A critical analysis of five examples* (pp. 1–176). New York, NY: Appleton.

Krantz, D. H. (1964). Conjoint measurement: The Luce-Tukey axiomatization and some extensions. *Journal of Mathematical Psychology*, *1*, 248–277. doi: 10.1016/0022-2496(64)90003-3

Krantz, D. H. (1971). Measurement structures and psychological laws. *Science*, *175*, 1427–1435. doi: 10.1126/science.175.4029.1427

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement: Additive and polynomial representations* (Vol. 1). New York, NY: Academic Press.

Krech, D. (1950). Dynamic systems, psychological fields, and hypothetical constructs. *Psychological Review*, *57*, 283–290. doi: 0.1037/h0062199

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, *355*, 584–585. doi: 10.1126/science.aal3618

Luce, R. D. (2000). *Utility of gains and losses: Measurement-theoretical and experimental approaches*. Mahwah, NJ: Erlbaum.

Luce, R. D. (2011, November 15). Inherent individual differences in utility. *Frontiers in Psychology*, *2*(297). doi: 10.3389/fpsyg.2011.00297

Luce, R. D., & Narens, L. (1994). Fifteen problems concerning the representational theory of measurement. In P. Humphreys (Ed.), *Patrick Suppes: Scientific philosopher* (Vol. 2, pp. 219–249). Dordrecht, the Netherlands: Kluwer Academic.

Luce, R. D., & Suppes, P. (2002). Representational measurement theory. In H. Pashler & J. Wixted (Eds.), *Stevens' handbook of experimental psychology* (3rd ed., Vol. 4, pp. 1–41). New York, NY: Wiley.

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

Mach, E. (1868). Über die Definition der Masse [On the definition of mass]. *Carl's Repertorium für Experimental-Physik, für physikalische Technik, mathematische und astronomische Instrumentenkunde*, *4*, 355–359.

Mach, E. (1896). *Die Principien der Wärmelehre* [Principles of the theory of heat]. Leipzig, Germany: Barth.

Maraun, M. D., & Halpin, P. F. (2008). Manifest and latent variables. *Measurement: Interdisciplinary Research & Perspective*, *6*, 113–117. doi: 10.1080/15366360802035596

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist*, *70*, 487–498. doi: 10.1037/a0039400

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

Michell, J. (1990). *An introduction to the logic of psychological measurement*. Hillsdale, NJ: Erlbaum.

Michell, J. (1994). Measuring dimensions of belief by unidimensional unfolding. *Journal of Mathematical Psychology*, *38*, 244–273. doi: 10.1006/jmps.1994.1016

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. doi: 10.1177/0959354300105004

Michell, J. (2008). Is psychometrics pathological science? *Measurement*, *6*, 7–24. doi: 10.1080/15366360802035489

Michell, J. (2017). On substandard substantive theory and axing axioms of measurement: A response to Humphry. *Theory & Psychology*, *27*, 419–425. doi: 10.1177/0959354317706746

Middleton, W. E. K. (1966). *A history of the thermometer and its use in meteorology*. Baltimore, MD: Johns Hopkins Press.

Mill, J. S. (1843). *A system of logic: Ratiocinative and inductive* (Vol. 1). London, UK: John W. Parker.

Morris, S. D., Grice, J. W., & Cox, R. A. (2016). Scale imposition as quantitative alchemy: Studies on the transitivity of neuroticism ratings. *Basic and Applied Social Psychology*, *39*, 1–18. doi: 10.1080/01973533.2016.1256288

Müller, T., Zimmermann, B., Meineke, J., Brantut, J.-P., Esslinger, T., & Moritz, H. (2010). Local observation of antibunching in a trapped Fermi gas. *Physical Review Letters*, *105*(4). doi: 10.1103/PhysRevLett.105.040401

Narens, L., & Luce, R. D. (1986). Measurement: The theory of numerical assignments. *Psychological Bulletin*, *99*, 166–180. doi: 10.1037/0033-2909.99.2.166

National Physical Laboratory. (1998). *Guide to the measurement of force*. London, UK: Institute of Measurement and Control.

Newton, I. (1999). *Mathematical principles of natural philosophy*. Berkeley: University of California Press. (Original work published 1687)

Open Science Collaboration. (2015, August 28). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. doi: 10.1126/science.aac4716

Orth, B. (1974). *Einführung in die Theorie des Messens* [Introduction to measurement theory]. Stuttgart, Germany: Kohlhammer.

Ostwald, W. (Ed.). (1894). *Das Ausdehnungsgesetz der Gase: Abhandlungen von Gay-Lussac, Dalton, Dulong und Petit, Rudberg, Magnus, Regnault* [The law of the expansion of gases: Treatises by Gay-Lussac, Dalton, Dulong and Petit, Rudberg, Magnus, Regnault]. Leipzig, Germany: Engelmann.

Pfanzagl, J. (1968). *Theory of measurement*. Würzburg, Germany: Physica-Verlag.

Poncet, S., & Dahlberg, L. (2011). The legacy of Henri Victor Regnault in the arts and sciences. *International Journal of Arts and Sciences*, *4*, 377–400.

Popper, K. (2002). *The logic of scientific discovery*. London, UK: Routledge. (Original work published 1935)

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Nielsen & Lydiche.

Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. *The Danish Yearbook of Philosophy*, *14*, 58–93.

Regnault, V. (1847). Relation des expériences … pour determiner les principales lois physiques et les données numériques qui entrent dans le calcul des machines à vapeur [Report on experiments … to determine the main physical laws and numerical data used in the calculation of steam engines]. *Mémoires de l'Académie Royale des Sciences de l'Institut de France*, *21*, 1–748.

Rettler, H. (1993). *Probleme der Metrisierung und Messung psychischer Merkmale* [Problems of metrization and measurement of psychological attributes]. Braunschweig, Germany: Technische Universität Braunschweig.

Roediger, H. L. (2008). Relativity of remembering: Why the laws of memory vanished. *Annual Review of Psychology*, *59*, 225–254. doi: 10.1146/annurev.psych.57.102904.190139

Rowlinson, J. S. (2010). James Joule, William Thomson and the concept of a perfect gas. *Notes and Records of the Royal Society*, *64*, 43–57. doi: 10.1098/rsnr.2009.0038

Sanner, C., Su, E. J., Keshet, A., Gommers, R., Shin, Y., Huang, W., & Ketterle, W. (2010). Suppression of density fluctuations in a quantum degenerate Fermi gas. *Physical Review Letters*, *105*(4). doi:10.1103/PhysRevLett.105.040402

Scott, D., & Suppes, P. (1958). Foundational aspects of theories of measurement. *Journal of Symbolic Logic*, *23*, 113-128. doi: 10.2307/2964389

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

Sixtl, F. (1982). *Messmethoden der Psychologie: Theoretische Grundlagen und Probleme* [Measurement methods in psychology: Theoretical foundations and problems] (2nd ed.). Weinheim, Germany: Beltz.

Stenner, A. J., Fisher, W. P., Jr., Stone, M. H., & Burdick, D. S. (2013, August 23). Causal Rasch models. *Frontiers in Psychology*, *4*(536). doi: 10.3389/fpsyg.2013.00536

Stepin, V. S. (2005). *Theoretical knowledge*. Dordrecht, the Netherlands: Springer.
Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*, 667–680.
Suppes, P. (1951). A set of independent axioms for extensive quantities. *Portugaliae Mathematica*, *10*, 163–172.
Suppes, P., & Zinnes, J. L. (1963). Basic measurement theory. In R. D. Luce, R. R. Bush, & E. H. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 1, pp. 3–76). New York, NY: Wiley.
Tarski, A. (1954). Contributions to the theory of models. I. *Indagationes Mathematicae*, *57*, 572–581. doi: 10.1016/S1385-7258(54)50074-0
Tolman, E. C. (1938). The determiners of behavior at a choice point. *Psychological Review*, *45*, 1–41. doi: 10.1037/h0062733
Torgerson, W. S. (1958). *Theory and methods of scaling*. New York, NY: Wiley.
Touré-Tillery, M., & Fishbach, A. (2014). How to measure motivation: A guide for the experimental social psychologist. *Social and Personality Psychology Compass*, *8*, 328–341. doi: 10.1111/spc3.12110
Trendler, G. (2009). Measurement theory, psychology and the revolution that cannot happen. *Theory & Psychology*, *19*, 579–599. doi: 10.1177/0959354309341926
Trendler, G. (2013). Measurement in psychology: A case of *ignoramus et ignorabimus*? A rejoinder. *Theory & Psychology*, *23*, 591–615. doi: 10.1177/0959354313490451
Tulving, E. (2007). Are there 256 different kinds of memory? In J. S. Nairne (Ed.), *The foundations of remembering. Essays in honor of Henry L. Roediger III* (pp. 39–52). New York, NY: Psychological Press.
van Bavel, J. (2016, May 27). Why do so many studies fail to replicate? *The New York Times*. Retrieved from https://www.nytimes.com/2016/05/29/opinion/sunday/why-do-so-many-studies-fail-to-replicate.html
Vittori, O., & Mestitz, A. (1981). Calibration of the "Florentine little thermometer". *Endeavor*, *5*, 113–118. doi: 10.1016/0160-9327(81)90043-0
Watson, J. B. (1913). Psychology as the behaviorist views it. *Psychological Review*, *20*, 158–177. doi: 10.1037/h0074428
Wensel, H. T. (1941). Temperature. In C. O. Fairchild, J. D. Hardy, R. B. Sosman, & H. T. Wensel (Eds.), *Temperature: Its measurement and control in science and industry* (Vol. 1, pp. 3–23). New York, NY: Reinhold.
Whewell, W. (1840). *The philosophy of the inductive sciences, founded upon their history* (Vol. 1). London, UK: John W. Parker.
Whitworth, J. (1858). *Miscellaneous papers on mechanical subjects*. London, UK: Longman.
Wright, B. D. (1997). A history of social science measurement. *Educational Measurement*, *16*, 33–45. doi: 10.1111/j.1745-3992.1997.tb00606.x

Author biography

Günter Trendler has a degree in psychology from the University of Mannheim (Germany). Currently he is working as technical employee in the domain of plant design at a leading international industrial services provider. Previously he was employed as technician at a leading provider of technical solutions in mechanized tunneling. The present article is partly based on preparatory studies for a project investigating the establishment of temperature measurement at the end of the 19th century.