Conceptual issues in psychological measurement

Borsboom, D.

Citation for published version (APA):
4. SCALES

It may be that the task of the new psychometrics is impossible; that fundamental measures will never be constructed. If this is the case, then the truth must be faced that perhaps psychology can never be a science...
- Paul Kline, 1998

4.1 Introduction

In the 1930’s, the British Association for the Advancement of Science installed a number of its members with a most peculiar task: To decide whether or not there was such a thing as measurement in psychology. The commission, consisting of psychologists and physicists (among the latter was Norman Campbell, famous for his philosophical work on measurement), was unable to reach unanimous agreement. However, a majority of its members concluded that measurement in psychology was impossible; Campbell (cited in Narens & Luce, 1986, p. 186), for example, asked “why do not psychologists accept the natural and obvious conclusion that subjective measurements (...) cannot be the basis of measurement”. Similarly, Guild (cited in Reese, 1943, p. 6) stated that “to insist on calling these other processes [i.e., attempts at psychological measurement] measurement adds nothing to their actual significance, but merely debases the coinage of verbal intercourse. Measurement is not a term with some mysterious inherent meaning, part of which may be overlooked by the physicists and may be in course of discovery by psychologists”. For this reason, Guild concluded that using the term ‘measurement’ to cover quantitative practices in psychology “does not broaden its meaning but destroys it”. Reese (1943, p. 6) summarized the ultimate position of the commission: “They [the members of the commission] argue that psychologists must then do one of two things. They must either say that the logical requirements for measurement in physics, as laid down by the logicians and other experts in the field of measurement, do not hold for psychology, and then develop other principles that are logically sound; or they must admit that their attempts at measurement do not meet the criteria and both cease calling these manipulations by the word ‘measurement’ and stop treating the results obtained as if they were the products of true measurement”.

It would seem that the members of the commission anticipated that the al-
ternative 'logically sound' principles for 'true measurement' in psychology would probably never be discovered. But perhaps they did anticipate their report to have the desired impact in the sense that psychologists would finally recognize their errors, and would stop the unauthorized use of terms like measurement and quantity. Interestingly, exactly the opposite has happened: Psychologists have developed an alternative, but generally use the term 'measurement' to denote every procedure of assigning numbers except the logically 'correct' one. That is, the theory of fundamental measurement (the 'true' measurement theory to which Guild refers) has been extended in such a manner that 'logically sound' principles have become available for psychological measurement situations, primarily through the development of conjoint measurement structures (Luce & Tukey, 1964; Krantz, Suppes, Luce, & Tversky, 1971). Ironically, however, not a soul uses that theory in the practice of psychological measurement: Every year there appears an enormous number of books that have 'psychological measurement' in the title, but few of them even contain a reference to this work. The logical foundation for psychological measurement has thus become available, only to be neglected by its presumed audience – and psychologists have continued to use the term measurement for everything else.

The gist of what has been called the 'axiomatic' approach to measurement (Cliff, 1992), of which the theory of fundamental measurement can be considered a special case, is that measurement is an essentially representational activity, i.e., a process of assigning numbers in such a manner as to preserve basic qualitative relations observed in the world (Narens & Luce, 1986). The result of this activity is called a measurement scale. Psychologists are familiar with this concept mainly through Stevens' (1946) famous typology of 'levels of measurement' in terms of nominal, ordinal, interval, and ratio scales. The scale type is often deemed very important for determining what kind of statistics may be used, and in this manner it exerts considerable influence on the practice of data analysis in psychology (or, in any event, on the conscience of psychologists doing the analyses). The prescriptive aspect of scales has been the subject of enduring controversies between measurement theoreticians and statisticians (Lord, 1953; Stevens, 1968; Gaito, 1980; Townshend & Ashby, 1984; Michell, 1986; Velleman & Wilkinson, 1993), mainly because statisticians refuse to be told what is admissible and what not by what they seem to perceive as an utterly impractical theory (Lord, 1953; Gaito, 1980). However, apart from generating such controversies and acting on the psychologist's statistical conscience, scales and the associated theory of measurement have not entered mainstream psychology at all (Cliff, 1992).

This does not mean that nobody works with representationalism in psychology. The original developers of the theory, such as Luce, Suppes, and Narens, continue to work out the mathematical basis of measurement theory, joined by a group of researchers united in the Society for Mathematical Psychology. In a completely different corner of psychology, the advocates of Rasch measurement frequently allude to the fundamental measurement properties of the Rasch model; notable in this context are Wright (1997), Roskam (1984), and Bond & Fox (2001). Finally, at a more conceptual level Michell (1990; 1997) has attacked the common practice in psychology and psychometrics using a line of reasoning based on the axiomatic theory of measurement. His efforts have had impact on at least one psychometrician
4.1 Introduction

(Kline, 1998), and may well influence more. These researchers look to the future, and some of them seem to regard the coming of the "revolution that never happened" (Cliff, 1992) as the only road to a truly scientific psychology (Kline, 1998; Bond & Fox, 2001). Or, like Luce (1996, p. 95), they view such developments as simply "inevitable", so that "the only question is the speed with which they are carried out".

The axiomatic theory of measurement thus has a certain apologetic quality about it. It is also strongly normative, or even prescriptive, as is evidenced by terminology such as 'admissible transformations', and the idea that performing an inadmissible transformation destroys the 'meaningfulness' of conclusions based on the data (see Michell, 1986, for a discussion of this view). Now, methodology is in a sense always normative, but there is no approach in psychological measurement - not even in latent variables analysis - that so pertinently presents itself as the gatekeeper of rationality. Treatises based on the approach also insist on empirical testability of hypotheses in a manner that almost suggests that, if a hypothesis cannot be directly tested, it is meaningless, or at the very least suspect. For example, Michell (2000) has characterized the a priori assumption that psychological attributes are quantitative, which indeed is a strong metaphysical assumption in many latent variable models, as a methodological thought disorder, and this leads him to label the entire field of psychometrics as pathological. The reason for this disqualification seems to be that the hypothesis is not directly testable in commonly used models like the factor model. Those familiar with the philosophy of science may see a parallel with a historical movement that shared both the strong normativity, the desire to demarcate between meaningful and meaningless propositions, and the emphasis on the dangers of metaphysics - namely, the Vienna Circle. In this respect it is interesting to note that fundamental measurement theory originated in roughly the same period as logical positivism - a mere two years separate Campbell's (1920) *Physics: The elements* and Wittgenstein's (1922) *Tractatus Logico-Philosophicus*. There seems to be a certain similarity between, on the one hand, the divide between the empiricist, verificationist orientation of logical positivism and the robust realist, falsificationist philosophy of Popper (1959), and, on the other hand, the schism between representational measurement theory and the latent variables approach.

This chapter develops this intuition by inquiring into the status of the measurement scale, the central concept of representational measurement. This inquiry serves two purposes. First, in view of the critical commentaries of Michell (1990; 1999; 2000), Kline (1998), and the Rasch movement (Bond & Fox, 2001), it is important to scrutinize the axiomatic approach to measurement carefully - not only with respect to its alleged normative force, but also with respect to the philosophical ideas on which it is based. But second, the present chapter will add considerable clarification to the strong conclusions reached in the previous chapter, by showing what a truly empiricist theory of measurement looks like. For representational measurement theory, when compared to the latent variables approach, is almost devoid of metaphysics. It explicitly recognizes that measurement scales are constructions, and in fact builds upon this idea in a way that, it must be said, is consistent, elegant, and powerful. Therefore, the representational measurement approach introduces a sharp contrast, which brings out the realism inherent in latent variable models.
stronger than any argument could do by itself. However, the present investigation will show that the idea, that measurement is a representational activity, is unsatisfying on a number of counts. In fact, it is argued here that representationalism fails to address some crucial issues in psychological measurement. The relevance of representational measurement for psychological research is therefore concluded to be limited.

4.2 Three perspectives on measurement scales

Representational measurement theory is aimed at specifying the conditions necessary for the construction of an adequate representation of empirical relations in a numerical system. From a formal perspective, this is conceptualized in terms of a mapping of one set of relations into another. The resulting representation is considered adequate if it preserves the observed, empirical relations. Semantically, the interpretation of the measurement process is in terms of a reconstruction of the measurement process. For example, numerical operations are conceptualized as corresponding to empirical operations, even though no scientist ever carried out these operations in the manner described by the theory. From an ontological perspective, scales cannot be considered anything but a construction. It could, of course, be held that these scales have referents in reality, for example objective magnitudes. However, such a realist interpretation, if endorsed, is external to the model, in contrast to the inherent realism in latent variables analysis.

4.2.1 The formal stance

Syntax  Representational measurement theory constructs measurement as the mapping of objects and relations between objects from an empirical domain into a numerical domain. Both are characterized in terms of set theory (Scott & Suppes, 1958; Suppes & Zinnes, 1963). We imagine a set of objects, which is is denoted \( A \), and a set of \( n \) relations holding between these objects, denoted \( R_1, R_2, \ldots, R_n \). A relation between objects may, for example, be one of dominance between objects (e.g., John is larger than Jane), between objects and stimuli (e.g., John 'dominated' an IQ-test item by solving it), or between stimuli (e.g. item 1 is more difficult than item 2). It may also be one of proximity or similarity (e.g., John’s political orientation is more similar to Jane’s than to Peter’s), which may again be considered in terms of similarity between objects, between stimuli, or between objects and stimuli (Coombs, 1964). Still other relations may be based on preference orderings, as is common in subjective expected utility theory. Whatever the precise nature of the relations is taken to be, they are always taken to be purely qualitative (representationalism takes ‘larger than’ to be a qualitative comparison). Often, there is some operation that can be interpreted as ‘combining’ two objects to create a new one. This combining operation is denoted \( \oplus \). Sometimes this operation is empirical, such as laying two rods end-to-end to create a new one, and in this case we speak of extensive measurement. Such an empirical operation of combining is known as a concatenation operation. Campbell (1920) believed that fundamental measurement must be extensive, that is, there must exist an empirical concatenation operation,
and treated all other measurement as ‘derived’ from these fundamental measures. However, it was later shown that there are cases where representational measurement works without there being an empirical concatenation operation (Luce & Tukey, 1964; Krantz, Luce, Suppes, & Tversky, 1971).

Taken together, the set of objects, the relation between them, and the combining operation form what is called an empirical relational system which we will call $\mathcal{O}$, which may be read as a shorthand for ‘observed’. This system is denoted as $\mathcal{O} = (A, R, \oplus)$. Now it is the business of representationalism to construct, entirely on the basis of the observed relations between objects in the set and the combinations of these objects, a numerical representation which preserves the information in the empirical system. This basically comes down to assigning to each object in $A$ a number from some numerical domain $N$, to find a mathematical relation $S$ that represents the empirical relation $R$, and to find a mathematical operation $\ast$ that matches the combining operation $\oplus$. The resulting representational system, call it $\mathcal{R}$, a shorthand for ‘representation’, is then denoted $\mathcal{R} = (N, S, \ast)$. Because the representation preserves all the information that was present in the empirical system, the relation between these systems is one of homomorphism (it is not isomorphic because more than one of the elements in the empirical system may map to the same number in the representational system). The combination of $\mathcal{O}$ and $\mathcal{R}$ is called a measurement structure. Measurement, in the representationalist view, is thus essentially a homomorphic representation of objects in a numerical system.

Representational measurement is called axiomatic, because its main strategy is 1) to assume certain axioms to hold with respect to the objects and the relations among them, 2) to prove mathematically that, given these relations, a homomorphic representation is possible (this is done in a representation theorem), and 3) to show under which transformations of the scale values this homomorphism is preserved (this is done in a uniqueness theorem). The latter proof essentially characterizes the transformations under which the representation stays invariant. It can be interpreted in terms of automorphisms (Narens & Luce, 1986): This means that the uniqueness theorem states the class of transformations which may be used to map the representation into itself, in such a way that no information is lost. Uniqueness results form the basis for the well-known ‘levels of measurement’ introduced by Stevens (1946). If the structure of the representation is invariant under all one-one transformations, we have a nominal scale; if it is invariant under all monotonic transformations, we have an ordinal scale; if it is invariant under all linear transformations, we have an interval scale; and if it is invariant under all affine transformations, we have a ratio scale. These four scale types do not exhaust the possible scale types (Krantz, Luce, Suppes, & Tversky, 1971), but will do for the present exposition.

**Semantics** The semantics of representationalism vary somewhat depending on whether one considers extensive measurement, for which a concatenation operation exists, or other forms of measurement. In the extensive case, the semantics can be based on a rather concrete connection of the measurement process and the manipulation of the assigned numbers through the concatenation operation, which is
itself mapped into a numerical operation. In cases of measurement that are not characterized by concatenation, the semantics of the theory are limited to representation itself. Here, the discussion will be limited to extensive measurement and one particularly important nonextensive case, namely additive conjoint measurement.

**Extensive measurement** The semantics of representationalism, and especially of extensive fundamental measurement as envisioned by Campbell (1920), are exquisite. The typical example for which the construction of representational scales is illustrative is the measurement of length. In this case, one may consider a set of objects, say, people, to form the set $A$. Further, a qualitative relation can be constructed as ‘not noticeably longer than’, denoted by $\preceq$, where ‘Jane $\preceq$ John’ means ‘Jane is not noticeably longer than John’. Finally, a concatenation operation $\oplus$ is available, namely we can lay Jane and John head-to-feet and compare the resulting combined entity, ‘Jane$\oplus$John’ to other people, or concatenations of other people, in the set. This gives the empirical relational system $\mathcal{O} = \langle A, \preceq, \oplus \rangle$. Now, we can map the relations in the empirical relational system into a numerical system in such a manner that all relations, holding between the objects in the empirical set, continue to hold between the numbers representing these objects. So, if Jane is is not noticeably longer than John, then the number representing Jane must be smaller than or equal to the number representing John. We can, as is usual among representational measurement theorists as well as carpenters, construct the representation in the set of positive real numbers, $\mathbb{R}^+$, so that each person is represented by a number in this set. A common way to do this is by comparing an object to a unit of measurement, such as a centimeter, by counting the number of units that must be concatenated in order to match the object. This is done through the construction of a so-called standard sequence of equal intervals (Krantz, Luce, Suppes, & Tversky, 1971). A ruler with centimeter marks is an instantiation of a standard sequence. Further we choose the empirical relation $\preceq$ to be represented by the numerical relation $\leq$, and the concatenation operation $\oplus$ by the numerical operation $+$. Suppose that John is assigned the value $\phi(\text{John}) = 1.85$ in the meter scale, and Jane the value $\phi(\text{Jane}) = 1.75$, so that $\phi(\text{Jane}) \leq \phi(\text{John})$. Now a comparison between John and Jane, with the unaided eye, will reveal that Jane is not noticeably longer than John, i.e., $\text{Jane} \preceq \text{John}$. So, it is indeed the case that $\preceq$ does a good job of representing $\preceq$. The representation will hold for all people $a, b, \ldots$ in the set $A$, and the technical way of expressing this is to say that $a \preceq b$ if and only if $\phi(a) \leq \phi(b)$. Also, we will find that the value assigned to the combined object Jane$\oplus$John will be $\phi(\text{Jane} \oplus \text{John}) = 3.60$, which is equal to $\phi(\text{Jane}) + \phi(\text{John}) = 1.75 + 1.85 = 3.60$. The representation of $\oplus$ by $+$ is therefore also adequate. It can furthermore be shown that the representation preserves all relevant relations in the empirical system, such as transitivity (if Jane$\preceq$John, and John$\preceq$Peter, then Jane$\preceq$Peter).

Thus, the mappings of the objects in $A$ into numbers in $\mathbb{R}^+$, of $\preceq$ into $\leq$, and of $\oplus$ into $+$ have succeeded. Moreover, it can be proven that the scale is invariant up the the choice for a unit of measurement (this is to say that it does not matter whether we express someone’s height in centimeters or in meters, as long as we do this consistently). Thus, the scale is insensitive to transformations of the
form \( \phi'(a) = c\phi(a) \), where \( \phi(a) \) is the original scale value, \( c \) represents a change in unit of measurement, and \( \phi'(a) \) is the resulting transformed value. This means that, if John and Jane are measured in centimeters rather than meters (so that \( c = 100 \)), all relations will continue to hold. For example, \( \phi'(\text{Jane}) + \phi'(\text{John}) = 175 + 185 = 360 \) will continue to match \( \phi'(\text{Jane} \oplus \text{John}) = 360 \). However, if we use a centimeter instead of a meter and give each measured object a bonus length of 100 centimeters (so that we are in fact performing a linear transformation of the form \( \phi''(a) = 100 \times \phi(a) + 100 \)), the mapping is destroyed. For now we would get, for Jane separately, \( \phi''(\text{Jane}) = 100 \times 1.75 + 100 = 275 \), and, for John separately, \( \phi''(\text{John}) = 100 \times 1.85 + 100 = 285 \). So, the sum of their scale values equals 560. But the concatenated object Jane \( \oplus \) John, when measured with this bonus centimeter, would receive a scale value of \( \phi''(\text{Jane} \oplus \text{John}) = 100 \times \phi(\text{Jane} \oplus \text{John}) + 100 = 360 + 100 = 460 \). Thus, the mathematical operation + ceases to be an adequate representation of the empirical operation \( \oplus \). The scale values may be multiplied, but not translated, because this destroys the homomorphism between the empirical and numerical systems. This is one way of saying that the measurement of length is on a ratio scale.

Campbell (1920) held that measures that are extensive are the only genuine cases of fundamental measurement. However, Michell (2000; 2001) has noted the interesting fact that the German mathematician Hölder had already shown in 1901 that Campbell was incorrect; he had axiomatically proven that distance was quantitative without invoking a concatenation operation. Campbell and his contemporaries were apparently unaware of Hölder’s work (Michell, 2001), and fervently defended the thesis that measurement without concatenation was not really measurement at all. This was the (incorrect) basis of the critique of the commission installed by the British Association for the Advancement of Science; for in psychology, it is generally difficult to identify an empirical concatenation operation. What this would require is something like the following. Suppose that I were to administer an intelligence test to a number of people (objects). Suppose further that John scores 100, and Jane scores 120. Now if I could concatenate (combine) the objects (Jane and John) in a suitable way, and this combination were shown to produce a score of \( 100 + 120 = 220 \), and if this were true not only for John and Jane but for all combinations of people, then I would have shown that an empirical concatenation operation exists and matches the numerical operation of addition. In general, this will not work in psychological measurement. Whether this is important or not is questionable, given the fact that, for centuries, carpenters and tradespeople did quite well in measuring all kinds of things without being aware of the importance of a concatenation operation, and in fact measured many attributes for which no concatenation operation was known (such as temperature). Fortunately, nobody listened to the commission members, for it seems that if we had to wait for empirical concatenation operations to be identified before we started measuring, any attempt at constructing psychological measurement instruments would surely be nipped in the bud. Now, the fortunate development of measurement theory has been to reject the restrictive account of Campbell, and to replace it with a more liberal account. The unfortunate development has been that some theorists have elevated the resulting framework to the same normative level that was originally
occupied by Campbell's theory, thus stating that there cannot be measurement if there is no homomorphic representation.

**Conjoint measurement** Although the viewpoints of the commission of the British Association for the Advancement of Science were unreasonable, the discussion of psychological measurement that followed the publication of the commission's report was instrumental in the development of measurement theory. In fact, the mathematical psychologists that took up the challenge ended up with a formalization of measurement that was far more powerful than Campbell's own, and has perhaps even been more important for physics than for psychology. The response of psychologists started with the explicit articulation of representationalism by Stevens (1946). Stevens' representationalism leaned heavily towards operationalism, because he defined measurement as "the assignment of numerals according to rule", where the nature of the rule involved is left unspecified, and Stevens was quite clear that this can be any rule. So, in Stevens' version, measurement occurs more or less by fiat; consequently, it is meaningless to ask whether something is 'really' being measured, because the fact that numerals are assigned according to rule is the sole defining feature of measurement. There is neither a need nor a place for postulating attributes which are prior to the measurement operation, as is explicitly done in latent variable theory. Representationalism, as it developed in the work of Krantz, Luce, Suppes, & Tversky (1971), followed Stevens in dropping the concatenation operation, and also retained the idea that measurement theory is a theory about numerical assignments. However, not any rule of assignment will do, because the assignment rule used must preserve the empirical relations as laid down in the empirical relational system. In essence, this boils down to the fact that representationalism takes off when the empirical relational system is already known, and then views it as its task not to explain how this relational system came into being (which many would consider to be the goal of latent variable models), but to formulate the rules for numerical assignment that preserve the relations in the system.

The broadening of the semantics associated with representationalism, which was a direct result of dropping the demand for empirical concatenation operations, provided an opening for constructing psychological measurement systems. For in this more liberal approach, measurement is no longer seen as necessarily representing empirical operations; any representation that mirrors empirical relations will do, if it complies with the demand that it forms a homomorphic representation. This follows directly from Stevens' move, which for a large part consisted in drawing attention away from the manner in which measurements are obtained (i.e., through concatenation), and towards their relations-preserving character. It also avoids the pitfall of degrading into operationalism, however, because it is possible that the relational system originates from distinct modes of assignment for different parts of the system. This is important, for while it may be possible to concatenate rigid rods of manageable length, it is arguably difficult to concatenate objects to match interstellar distances, or to place Jupiter on a balance scale. Still, my Encyclopedia mentions the fact that the average distance between the earth and the sun is about 149597890 kilometers, and that the mass of Jupiter is approximately $1.967 \times 10^{27}$
kilograms; and I strongly suspect that the writers of my Encyclopedia mean these statements to refer to qualitatively the same dimensions as, say, the distance between my cup of coffee and my telephone, and the mass of the computer I am now working on. In the rigid version of measurement theory, which leads directly to Bridgman’s (1927) operationalism, these interpretations are not warranted; but in the more liberal representationalist interpretation, they are. Moreover, any imaginable structure that allows for a homomorphic representation can be subsumed under the general category of measurement. This includes structures observed in psychological measurement.

The class of structures most important to the present discussion is the class of additive conjoint structures (Luce & Tukey, 1964; Luce, Suppes, & Tversky, 1971; Narens & Luce, 1986). Additive conjoint structures pertain to relations between at least three variables. Two of these variables are considered ‘independent’ variables and one is ‘dependent’. The meaning of these terms is similar to that used in analysis of variance. What conjoint measurement does is a little strange from the psychometrician’s point of view, because the measurement relation is not defined on any of the three variables, but on all three simultaneously. Call the independent variables A and B, and the dependent variable Y; their levels are denoted a, b, and y, respectively. What is represented in conjoint measurement is the Cartesian product $A \times B$, which consists of all ordered pairs $(a, b)$, and the relation that is mapped in $\geq$ is the effect of these combinations on the dependent variable Y. Denote the levels of the independent variable A by $i, j, k$ and the levels of the independent variable B by $l, m, n$. The idea is that, if the joint effect of $(a_i, b_j)$ exceeds that of $(a_j, b_m)$, so that $(a_i, b_j) \geq (a_j, b_m)$, where $\geq$ again is a qualitative relation and not a quantitative one, then the combination $(a_i, b_j)$ must be assigned a higher number than the combination $(a_j, b_m)$. The process of quantification (i.e., representing qualitative relations in the real numbers) now applies to all three variables simultaneously, but it does not require an empirical concatenation operation. What happens is that the variables A, B, and Y are scaled at once through a quantitative representation of the trade-off between A and B in producing Y (Narens & Luce, 1986). The representation theorem for conjoint measurement axiomatically states the conditions under which this can be done. It turns out that the possibility of constructing a homomorphism hinges on the possibility to find a representation that is additive in the effects of the independent variables on the dependent variable. If this is the case, then mappings $f$ and $g$ of the independent variables A and B into the real numbers can be found so that $(a_i, b_l) \geq (a_j, b_m)$ if and only if $f(a_i) + g(b_l) \geq f(a_j) + g(b_m)$. The representational structure for the Cartesian product terms $(a, b)$ is for any combination of levels $i$ of A and $l$ of B then given by $\phi(a_i, b_l) = f(a_i) + g(b_l)$. The representation is on an interval scale, because the structure is invariant under linear transformations of the assigned scale values.

Conjoint measurement thus constructs a mapping of an empirical relational system $O = \langle A \times B, \geq \rangle$ into $R = (\mathbb{R}e, \geq)$. No empirical concatenation operation is in sight, as is reflected by the omission of $\oplus$ in the notation, although adding scale values is meaningful. It is possible to imagine a kind of concatenation operation that relates the system to extensive measurement structures, but this concatenation is not empirical. In the literature, this is called an ‘induced’ concatenation operation.
Scales

(Narens & Luce, 1986). Because this is not an empirical concatenation operation, however, the operation cannot be said to be represented in the way this can be said of extensive measurement.

It is, however, important to consider why conjoint measurement gives an interval scale, i.e., what the meaning of the measurement unit is. This comes down to the question what it is, exactly, that is being measured. Basically, what is represented in the model is a trade-off. The meaning of the measurement unit is in terms of this trade-off. For instance, suppose that we have a given combination \((a_i, b_l)\), and increase the level of \(A\) from \(a_i\) to \(a_j\), thereby constructing a new combination \((a_j, b_l)\) that is \(\succ\) to the original one. The conjoint model then says by how many units the factor \(B\) has to be decreased in order to produce a new combination \((a_j, b_k)\) that is not noticeably different from (i.e., that is both \(\succeq\) and \(\preceq\) to) the original combination \((a_i, b_l)\). Thus, the model states how effects resulting from variations in \(A\) can be undone by variations in \(B\), and vice versa. The measurement unit is explicitly defined in terms of this trade-off. The reason for this is that any two distances \(a_i - a_j\) and \(a_j - a_k\) on the factor \(A\) are defined to be equal if they can be matched by the same distance \(b_l - b_k\) on the factor \(B\). The measurement unit on the factor \(A\) is thus defined as the change in \(A\) necessary to match an arbitrary change on the factor \(B\), and the measurement unit on the factor \(B\) is defined as the change in \(B\) necessary to match an arbitrary change in the factor \(A\). This is the reason why it is crucial to have two factors; one cannot define a unit of measurement on one factor without reference to the other. Because the method does not match levels in \(A\) by levels in \(B\), but rather differences between levels of \(A\) by differences in levels of \(B\), it can be expected to yield an interval scale. This is formally the case because linear transformations on the scale values assigned to the levels of either of the factors preserve the representation.

4.2.2 The empirical stance

Representational measurement is, as has been stated before, concerned with formulating the conditions that must be fulfilled in order to be able to construct a representation. These conditions, which are formulated as axioms, thus describe the relations that must hold in the data at hand for a representation to be possible. They are of an empirical nature; in Krantz, Luce, Suppes, & Tversky (1971) they are even called empirical laws. For extensive measurement, the axioms involved are rather simple (see Narens & Luce, 1986, for a lucid description). For conjoint measurement, they are more complicated. Basically, if one knew a priori that the effects of the independent variables were additive, there would be no need for the specification of the axioms involved, and an additive representation could be readily constructed. The strategy of representationalism, however, is not to posit variables and relations between them in reality and to look at whether the data structure is more or less consistent with these (i.e., the model fitting approach as used in latent variable modeling). It always starts with the data, never with the metaphysics. So, the axioms of conjoint measurement describe characteristics that the data must exhibit for us to be able to construct an additive representation.

As always, what we start with is a set of purely qualitative relations. In this
4.2 Three perspectives on measurement scales

In this context, however, the elements on which these relations are defined are the combinations \((a, b)\). These combinations are assumed to be ordered. This ordering is in a sense ‘induced’ by \(Y\). For example, suppose that a subject must judge, for a tone generated by a given combination \((a, b)\) of intensity \((A)\) and frequency \((B)\) whether its loudness \((Y)\) noticeably exceeds \((\geq)\) that of a tone generated by a different combination. The first axiom of conjoint measurement states that the ordering so produced must be a weak order. A weak order is an order that is transitive and connected. Transitivity means that for each combination of levels \(i, j, k\) of \(A\) and \(l, m, n\) of \(B\), if \((a_i, b_l) \geq (a_j, b_m)\) and \((a_j, b_m) \geq (a_k, b_n)\), then \((a_i, b_l) \geq (a_k, b_n)\). Connectedness means that each comparison is made, and for all comparisons either \((a_i, b_l) \geq (a_j, b_m)\), or \((a_j, b_m) \geq (a_i, b_l)\), or both.

The second axiom of conjoint measurement is called independence. It states that the ordering of the levels in \(A\), which is induced by the ordering in \(Y\), must be unaffected by which particular value of \(B\) is chosen to assess this ordering; the converse must also hold. So, if we assess the ordering of perceived loudness as produced by varying levels of intensity, we have to do this while holding the frequency of the presented tones constant. The independence condition says that it must not make a difference for the ordering whether we set the frequency at 100Hz or at 1000Hz. Higher intensities must in either case produce either an unnoticeable difference or a higher perceived loudness. This means that, if there is an interaction effect of the independent variables, no additive conjoint measurement representation can be formed. However, the restriction this poses is less serious than it may seem. This is because the original observations on the \(Y\) variable are assumed to be merely ordinal. Thus, any monotonic, order-preserving transformation on these observations is permissible. The restriction posed is therefore relatively mild: There must exist a monotonic transformation of the dependent variable that renders the effects of the independent variables additive. It is possible to remove a wide class of interaction effects by transforming the dependent variable. A real problem occurs, however, in the presence of disordinal interactions, i.e., when effects ‘cross’. This would be the case, for example, if for tones with a frequency below 1000Hz a higher amplitude would produce a higher perceived loudness, but for tones with a frequency above 1000Hz, a higher amplitude would produce a lower perceived loudness. If this happens, the very ordering on \(A\), as induced by the ordering on \(Y\), depends on the selected level of \(B\), and no additive representation will be possible.

The independence condition allows for the independent ordering of the factors \(A\) and \(B\) in terms of increasing values of \(Y\). On the basis of this ordering, we can represent the structure in a table like Table 1, which contains three levels for each factor. Factor \(A\) is represented as increasing in \(Y\) from left to right; factor \(B\) is represented as increasing from top to bottom. The entries in the table are the (monotonically transformed) values \(y\) as corresponding to each combination \((a, b)\). Because of the independence condition, the entries are increasing both in the rows and in the columns of the table.
Table 4.1. The combinations \((a,b)\) are ascending both in rows (left to right) and columns (top to bottom).

<table>
<thead>
<tr>
<th>Factor A</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>((a_1, b_1))</td>
<td>((a_2, b_1))</td>
<td>((a_3, b_1))</td>
</tr>
<tr>
<td>Factor B</td>
<td>2</td>
<td>((a_1, b_2))</td>
<td>((a_2, b_2))</td>
</tr>
<tr>
<td></td>
<td>((a_1, b_3))</td>
<td>((a_2, b_3))</td>
<td>((a_3, b_3))</td>
</tr>
</tbody>
</table>

The third axiom of conjoint measurement is called double cancellation and refers to relations between the diagonals of the table. It is basically a consequence of additivity, and invoked because the axiom of monotonicity does not, by itself, guarantee additivity in the two-factor case (Krantz, Luce, Suppes, & Tversky, 1971, p. 250). Additivity requires that any entry \((a, b)\) can be represented by the additive function \(f(a) + g(b)\). Therefore, an entry, say, \((a_2, b_1)\), must be \(\geq\) (yield a greater amount of \(Y\)) to another entry, say, \((a_1, b_2)\), if and only if \(f(a_2) + g(b_1) \geq f(a_1) + g(b_2)\). Suppose that this is the case, and that it is also the case that \((a_3, b_2) \geq (a_2, b_3)\). Then we have the two inequalities

\[
f(a_2) + g(b_1) \geq f(a_1) + g(b_2) \quad (4.1)
\]

and

\[
f(a_3) + g(b_2) \geq f(a_2) + g(b_3). \quad (4.2)
\]

If the effects of the factors are additive, it follows that

\[
f(a_2) + g(b_1) + f(a_3) + g(b_2) \geq f(a_1) + g(b_2) + f(a_2) + g(b_3), \quad (4.3)
\]

which implies the new inequality

\[
f(a_3) + g(b_1) \geq f(a_1) + g(b_3). \quad (4.4)
\]

This is the condition of double cancellation ("cancellation", because of the terms cancelling out in the last step of the derivation, and "double" because there are two antecedent inequalities). The double cancellation axiom must hold for all \(3 \times 3\) submatrices of the larger matrix defined over all levels of \(A\) and \(B\).

The final axiom needed is called the Archimedean axiom. This axiom is also commonly used in extensive measurement, where it asserts, for instance, that no object is infinitely larger than any other object. In the present context, the axiom states that no difference in \(A\) produces an infinitely larger change in \(Y\) than any other difference in \(A\), and that no difference in \(B\) produces an infinitely larger change than any other difference in \(B\). This axiom is technical in nature, and I will neglect in the following.

If the data satisfy the above axioms, then an additive representation can be constructed that preserves all of the relevant relations in the data. Conjoint measurement theory shows that fundamental measurement does not require a concatenation operation, and in doing so provides a justification for intensive measurement that is lacking in Campbell's account. It also provides psychology with a system for
measurement that is on equal footing with the ones in physics. For the way subjective loudness could be measured and quantified is exactly the same way in which density can be measured and quantified. The representationalists thus showed that the conclusion reached by the commission installed by the British Association for the Advancement of Science was false: Fundamental measurement is, in principle, possible in psychology.

4.2.3 The ontological stance

Representationalism is the only theory of measurement with an explicit ontological status for its central concept. Scales are representations of observed relations and therefore they are constructions. Scales do not ‘underlie’ the observed relations, and much less are they causally active in producing them. The recipe for scale construction is crisp and clean, and devoid of any metaphysical assumptions whatsoever. Parsimonious and powerful, representationalism is the dream of every empiricist philosopher and scientist alike. We can simply start by observing relations between objects with the unaided eye (Van Fraassen, 1980), and show how theoretical terms like ‘length’, ‘distance’, or ‘subjective loudness’ can be constructed and quantified based on these relations. There is little one can say about this except that it is probably the most comprehensive and adequate empiricist theory of measurement that could possibly be given. Upon closer consideration, however, it appears to be rather unclear what exactly constitutes the relation between the logical structure of representationalism and the actual measurement process. One option is to interpret the axiomatic theory as providing us with a definition of measurement. This would suggest that the theory provides necessary and sufficient, or at the very least necessary, conditions for a procedure to satisfy in order to be covered by the definition of measurement. This interpretation will be argued to be problematic; if the theory gave sufficient conditions, then it would include absurd cases, and if it gave necessary conditions, many recognized instances of measurement would not be covered by the definition because representationalism cannot deal with error. Alternatively, one could interpret the theory as a prescriptive theory that shows how we should go about constructing scales in psychology. This, however, does not work either. Contrary to suggestive wordings like ‘not noticeably longer than’, representationalism does not describe how fallible human beings such as ourselves could construct scales. It describes, at best, how a Laplacean demon – a rational being with an infinite amount of time, an infinitely large brain, and a capacity for errorless observation – could construct scales on the basis of observable qualitative relations. Because we are not such beings, the representationalist enterprise cannot be seen as a serious proposal for constructing measurements. Therefore, the prescriptive reading of the theory is not justified.

Representation and measurement

The problem with the view that representationalism gives a definition of measurement concerns its central tenet, namely that measurement is essentially about representation. While there is a nontrivial sense in which this is true, namely, we do
aim to construct a numerical system that reflects certain systematic relations in the world, there is also a nontrivial sense in which it is false. The sense in which it is false is that measurement is not exclusively representational. In particular, the fact that a representation can be constructed cannot be a sufficient condition in any sensible definition of measurement.

This is evident from the fact that we can construct situations where we have homomorphic representations which are not measurements in any meaningful sense of the word. Consider, for example, the Guttman model, which is a deterministic item response model to which the axiomatic theory applies. The Guttman model is generally seen as a model for ordinal measurement, but its mathematical requirements by themselves do not warrant this interpretation. To see this, consider the following four items:

1. I have parents (yes: 1, no: 0)
2. I have no beard (yes: 1, no: 0)
3. I menstruate (yes: 1, no: 0)
4. I have given birth to children (yes: 1, no: 0)

Suppose that we administered these items to a group of people. Obviously, we would get a triangulated structure that looks as follows:

<table>
<thead>
<tr>
<th>Item 1</th>
<th>Item 2</th>
<th>Item 3</th>
<th>Item 4</th>
<th>Sumscore</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
</tbody>
</table>

This triangulated structure is a necessary and sufficient condition for constructing a Guttman scale. The reason that we get this structure, of course, is simply that we have constructed inclusive subclasses of people. People with sumscore 1 are men with a beard; people with sumscore 2 are non-menstruating women and men without a beard, people with sumscore 3 are women without children, and people with sumscore 4 are women with children. Now, if measurement were nothing more than homomorphic representation of empirically observed relations, and the Guttman model produces an ordinal scale, then we would be forced to conclude that we have ordinally measured something here. This does not seem to be the case. However, the example surely provides a case of homomorphic representation. Therefore, representation and measurement are not the same. That a representation can be constructed is not a sufficient condition for obtaining measurements.

This is not surprising because a representation is a purely formal concept, while the question whether measurement has taken place is not a purely formal one, as is evident from the literature on validity (Cronbach & Meehl, 1955; Messick, 1989). Thus, to speak of measurement requires extending the formal framework with a substantive interpretation. And this interpretation cannot, in principle, be given by the formal model itself. It seems to me perfectly in order to say that a
representation may be constructed in cases where nothing is measured. Now this is not an argument against the representational view in general; it merely says that there is more to measurement that representation alone, and that the ability to construct a representation cannot be a sufficient condition for measurement.

**The problem of error**

While it is quite clear that representationalism cannot give sufficient conditions for measurement, we could at least imagine that the theory gives necessary conditions for measurement, as is suggested, for example, by Michell (1990; 1999). This, however, is also difficult because the theory has a hard time dealing with the problem of error. If the possibility to construct a homomorphic representation were to be a necessary condition for measurement, this entails that we should be able to gather data that fit the measurement model perfectly. This is because, strictly speaking, models like the conjoint model are refuted by a single violation of the axioms. For example, if there is a single triple of observations where transitivity is violated, or a single $3 \times 3$ submatrix that violates double cancellation, the model is falsified, because no homomorphic representation will be possible. Since we can safely assume that we will not succeed in getting error-free data – certainly not in psychology – we must choose between two conclusions: Either measurement is impossible, or it is not necessary to construct a perfect homomorphic representation. If we accept the former, we may just as well stop the discussion right now. If we accept the latter, then we have to invent a way to deal with error.

**The return of Mr. Brown**

The natural means of introducing a theory of error would be to construct a statistical formulation of representational measurement theory. In such a theory, one would have to introduce parameters to represent the true values of the objects. One way to do this would be by replacing sentences like `$a \preceq b$ if and only if $\phi(a) \leq \phi(b)$' with sentences of the form `$\tau_a \preceq \tau_b$ if and only if $\phi(a) \leq \phi(b)$'. Here, the $\tau$ variable could serve the function of denoting the true value of the objects on some instrument used to make the comparison between $a$ and $b$. This instrument could be a particular meter stick, but it could also be an item in a psychological test. Scheiblechner (1999) who follows this line of reasoning, calls the indirect comparison of objects, through their true scores on an instrument, an instrumental comparison (p.299). The so constructed model allows for error because it may be the case that a particular observer judges that $a < b$ while it is actually the case that $\tau_a > \tau_b$.

The problem, of course, is that the very introduction of error requires an account of what the true values are. The common approach to this problem in statistics is by introducing the idea that the observed values are realizations of a random variable. Conceiving of the measurement apparatus as yielding a value $x$ for each object, we could implement this idea by interpreting $x$ as a realization of the random variable $X$. We may then introduce the assumption that $\mathcal{E}(X_a) = \tau_a$, analogous to the way this is done in classical test theory. The interpretation of the so constructed sentence in terms of length would be 'the expected centimeter reading of $a$ is not noticeably larger than the expected centimeter reading of $b$ if and only if the number
assigned to \( a \) is smaller than the number assigned to \( b \)'. Because nobody can observe the expected values, we should delete the word 'noticeably'. This implies that we should also replace the symbol \( \leq \), which stands for 'not noticeably longer than' by the symbol \( \leq \) which means 'has a lower expected centimeter reading than'\(^1\). That is, the instrumental comparison can only be made by examining relations between expected values, which are by necessity numerical. So, an interesting shift takes place here: While the fundamental measurement model aims to construct quantitative metrics from qualitative observations, the instrumental comparison introduces a kind of quantitative metric directly at the level of the comparisons made.

Expected values are not observable, and the fact that we are introducing relations between unobservables at such a fundamental level in the construction of the model has far-reaching consequences. In effect, we are now already working with a true score model. And if we aim to construct a measurement instrument that measures a single attribute with a number of observed variables, we will build a structure that strongly resembles a latent variable model. Considered in terms of a psychological test consisting of a number of items, this would work as follows. Interpreting the numerical assignment \( \phi \) as a latent variable (now interpreted as a rescaling of the true score), which represents an item \( \times \) subject Cartesian product with an ordering induced (in both items and subjects) by the \( \tau \) variable, we can construct an additive conjoint representation if the item and subject effects are independent, additive, and satisfy the double cancellation axiom with respect to the values of \( \tau \) (Scheiblechner, 1999). An example of a model that has these properties is the Rasch model (Rasch, 1960). Thus, this statistical decomposition of observed values, in true and error components, leads directly to the class of additive Item Response Theory models. I will have more to say about this relation in Chapter 5.

This approach to the problem of error is useful because it shows that the di-vide between representationalism and latent variable theory is formally speaking a fine line. From a philosophical viewpoint, however, crossing this line has serious consequences; in effect, the main tenets of representationalism are lost in the present approach. The first problem is that we have assumed the existence of an instrument that gives the measurements to apply the expectation operator to. The present approach merely allows for the construction of a ruler with equal intervals on the basis of comparisons made by using a ruler with unequal intervals. It can be used to show how a scale can be linearized, analogous to the way that Rasch models linearize sumscores by appropriately stretching the far ends of the scale. However, representationalism is not served by assuming, a priori, that a ruler exists. For the theory is aimed at showing how a ruler with equal intervals could be constructed on the basis of direct qualitative comparisons with respect to the variable in question.

---

\(^1\) While probably not intended in this manner, the fact that Scheiblechner (1999) retains the \( \leq \) relation in the introduction of his ADISOP models is slightly misleading. The relation is interpreted as a 'stochastic dominance' relation (p. 299), where \( a \leq b \) means, for example, that subject \( a \) has a lower expected value on a given item than subject \( b \). The notation is adequate in that the denoted relation is only taken to establish an ordering, and in this sense is qualitative, but it does certainly not have the connotation of noticeability or observability, a connotation that such relations generally do have in the representational approach.
– whether it is length, density, or subjective loudness – and not at showing how length can be measured given the fact that a ruler already exists. More importantly, however, the very construction of a ruler is frustrated in the present approach. The reason for this is that the construction process would have to be carried out though the evaluation of stochastic dominance relations of the above type. These relations are clearly unobservable. Moreover, expectations cannot be empirically concatenated in principle. As a result, even the possibility of extensive measurement now vanishes. The third and most troubling consequence of this move is that in most cases of measurement, but certainly in psychology, we will encounter serious problems in the interpretation of the expected values involved. In fact, we are likely to be forced to interpret the expected values in a propensity sense. So we can now hear Mr. Brown knocking on the back door; and the representationalist certainly would not want to let him in. It thus seems that, in this approach, we are quickly losing the gist of the representationalist theory. For we are not building homomorphic representations of observable objects and qualitative observable relations between them; we are building isomorphic representations of unobservable true scores and equally unobservable relations between them.

**Introducing the Laplacean demon**  
A second way to introduce a concept of error would be to introduce true relations between objects, rather than to assume true scores for the objects. This could be done by replacing sentences like \(a \leq b\) if and only if \(\phi(a) \leq \phi(b)\) with sentences of the form \(a \preceq_{\text{true}} b\) if and only if \(\phi_{\text{true}}(a) \leq \phi_{\text{true}}(b)\). That this will not work is obvious from the fact that the values \(\phi\) are, in representationalism, constructed from the data and not hypothesized a priori. Because we cannot observe the true relations, we cannot construct these values and the above formulation is nonsensical. It would be an idea, however, to take the idealization one step further and to introduce true values for the \(\phi\) involved. These values are not to be interpreted as existing independently of the relations they figure in, as in the introduction of expected values above. Rather, they should be seen as the values that would be constructed if we could observe the true relations between objects. Their status as true values is thus derived from positing true relations, rather than the other way around. Also, the relation \(\preceq\) does not have to be interpreted as a relation between propensities. It can be taken to be a completely deterministic relation between objects. So now we could get \(a \preceq_{\text{true}} b\) if and only if \(\phi_{\text{true}}(a) \leq \phi_{\text{true}}(b)\). Interpreted in terms of length, this sentence says that \(a\) is truly not noticeably longer than \(b\) if and only if the number, that would be assigned to \(a\) if we could observe the true relations, is smaller than or equal to the number, that would be assigned to \(b\), if we could observe the true relations. We thus retain the construction of quantitative scales out of qualitative relations, and refrain from introducing relations between unobservables in the definitions. The only problem is that the relation \(\preceq\) has no natural interpretation anymore. For what does ‘truly not noticeably longer than’ mean? Does it mean that nobody could, in principle, notice that \(a\) is longer than \(b\) if \(a\) is actually longer than \(b\)? No, because if this were the case, we could just use fundamental measurement theory as it stands; for there would be no error, and consequently there would be no need for the present
exercise. Does it then mean that no perfect observer could notice that \( a \) is longer than \( b \) if \( a \) is truly longer than \( b \)? Possibly, but who is the perfect observer? A Laplacean demon?

The problem we are facing here is clearly caused by the word ‘noticeably’. The use of this term suggests that somebody is actually noticing relations between objects, and the judgments of this anonymous somebody would produce transitive and ordered data when measuring attributes that sustain measurement. Upon closer inspection, the identity of this anonymous observer is mysterious. The interpretation of the word ‘noticeable’ is unproblematic for an empiricist reading of the theory as long as we interpret it as ‘noticeable with the unaided eye’, that is, noticeable in practice. Because in this interpretation the theory is unable to deal with error, we have to move beyond the practically feasible observational powers of human beings and construct the relations as noticeable for somebody with observational powers that markedly exceed our own. This is necessary because the introduction of error means that we need to be able to say that we are wrong, and being wrong is always relative to being right. That is, error is a deviation, and the natural interpretation of the concept is that it is a deviation from a standard. In a theory that works its way up from qualitative, noticeable relations, we need somebody to notice the correct relations, which could function as such a standard. And if it cannot be us, then it must be a demon with infallible observational powers. Hence the need to introduce a Laplacean demon.

Now, if we want to pursue this line of reasoning without introducing propensities, expected values, and latent variables into reality, it is obvious that we must limit the relation \( \preceq \) to be a relation between objects, and not between true scores. If we do not do this, then we must again introduce expected values and relations between them for the demon to notice. This requires that such values and relations exist in reality, so that we would again be introducing the metaphysics we sought to evade; in effect, we would arrive at the same conception of measurement as in the previous attempt to deal with error. Dismissing relations between propensities, however, has a very important consequence: It excludes any model that posits relations between expected values. Thus, in this interpretation, additive models like the Rasch model (1960) and the ADISOP models (Scheiblechner, 1999) are not representational models because they posit relations between propensities.

Perhaps the representationalist would not object to the exclusion of additive IRT models. One rarely encounters a reference to these models in the representationalist literature, and I would indeed suspect that representationalists reject such models because of the fact that they introduce too much metaphysics. The advocates of additive IRT models tend to flirt with fundamental measurement theory (e.g., Wright, 1997; Bond & Fox, 2001), but the reverse is definitely not the case. However, even the pet examples of representationalism would have difficulty surviving the demands posited in the approach we are presently considering. Consider the measurement of subjective loudness. What would we have to posit in order to be able to say that, while subject \( i \) did not notice the combination \((a_i, b_i)\) to be \( \preceq \) to the combination \((a_j, b_k)\), he erred in this response? Or to say that, while \( i \) said he preferred stimulus \( j \) to stimulus \( k \), he was misjudging his own preferences? The problem here is, of course, that the word ‘noticeable’ is, in these cases, intended as
'noticeable for subject i' and not as 'noticeable for a Laplacean demon'. The very subjective nature of the comparisons on which fundamental measurement operates in these cases precludes the introduction of error. For this requires us to say that i is objectively wrong concerning his subjective state of mind. This does not seem to go in the right direction, whether we consider the situation from the representational point of view or otherwise. Thus, in this approach few of the accomplishments of representational theory are preserved: The additive IRT models are excluded from consideration, and subjective scales for loudness, preference, etc., are deprived of their intended interpretation. In fact, the only examples of measurement that would sit well with this approach are examples from physics. The measurement of length, mass, and density do sustain the idea that they represent deterministic relations between objects as they could be observed by a Laplacean demon. But the measurement of psychological variables is not satisfactorily incorporated in this approach.

Reconceptualizing error  A final possibility to deal with imperfect observations is not to view them as error at all. Whatever the ontological status of error may be, in the final analysis the only epistemological criterion to detect error is as a deviation from a theoretical model. Instead of laying the blame on the observations, so to speak, one may attribute the deviations to a failure of the model. In such a view, the model is not interpreted as aspiring truth, but as an approximation. One may then choose to minimize the distance between, for instance, the conjoint representation and the data matrix. This can be done by constructing a stress measure for this distance, and then minimizing the stress of the model with respect to the data. Interpreted in this manner, representational measurement theory would be a (possibly multidimensional) scaling technique, because error is not conceptualized as inherent to the variables observed, but as the distance between the data matrix and the representation. (In multidimensional scaling it is nonsensical to ask what the 'true' representation is, in contrast to latent variable models, where the quest for the true model is often deemed very important.) Representationalism does have a structure that is similar to scaling techniques (Coombs, 1964), so that this approach would seem a natural way for representationalism to deal with error. However, in this approach the main idea of representational measurement theory is also lost, because whatever the relation between the data and the representation may be, it will not be a homomorphic mapping.

So, it seems that representational theory is stuck between a rock and a hard place: It must either say that no psychological set of data satisfies the axioms, thereby forcing the conclusion that psychological measurement is impossible after all, or it must introduce a concept of error. The three ways of doing this, as discussed above, are not satisfactory. In the first attempt, we were forced to introduce expected values for the objects. This not only requires the existence of an instrument yielding values to apply the expectation operator to, but must also posit probabilities that can only be interpreted as propensities. In effect, the structure we end up with strongly resembles a latent variable model, and the homomorphism constructed involves unobservable relations between unobservable true scores. This
can hardly be considered to maintain the spirit of representationalism. The sec­ond attempt introduced true qualitative relations between the objects, and derived true values only in virtue of these relations. However, in this interpretation we must hypothesize a supernatural being to observe the true relations. Although this conception is perhaps closest in spirit to the formulation of representational measurement, it cannot be considered a case of progress in terms of keeping a low metaphysical profile. Finally, if we choose a more pragmatic approach, and simply minimize the distance between the data and the representation, we refrain from introducing metaphysics, but at the same time lose another central idea in representationalism, which is that we are constructing a homomorphic mapping of objects into the real numbers. Thus, the inability to deal with error seems to be deeply entrenched in the structure of representationalism. Attempts to incorporate an error structure seem to invariably destroy one or another of the tenets of the theory. This does not, of course, imply that the formal structure of representational theory could not be applied to stochastic systems. It merely means that to do so requires giving up the empiricist connotation of the theory.

Representationalism as rational reconstruction

Representationalism does not state sufficient conditions for a definition of measurement, because there are representations that are not measurements. Neither does it provide necessary conditions, because the conditions as stated will be false in the presence of error, and it is hard or impossible to modify the theory to bypass this problem. We cannot hold, therefore, that representationalism defines measurement in general. We may, at best, hold that it states necessary conditions for perfectly reliable (i.e., deterministic, errorless) measurement, which would be a definition without much practical use – particularly in psychology. If it does not offer a definition of measurement, however, what is the status of representationalism? What is its relation to actual measurement?

First, we must note that representational theory certainly elucidates the structure of measurement. Especially in physics, the theory has greatly clarified the nature of various measurement techniques by concentrating on the relation between a empirical relational system and a numerical relational system. Second, it is clear that the theory seeks to offer logical underpinnings of scale construction. In the extensive case, the logical requirements are connected to the process of measurement by relating the concatenation operation to the numerical operation of addition. This is a particularly interesting case because it seems to show what it ‘actually is’ that carpenters and tradespeople have been doing all along. The theory gives a formal structure, and it would seem that this formal structure is in some way ‘instantiated’ in the behavior of cashiers and scientists alike. This suggests that the theory gives a reconstruction of the measurement process.

Of course, such a reconstruction should definitely not be taken to be an actual reconstruction of the historical process that led to measurement as we now know it. Such an interpretation of the approach would be vulnerable to the same objection that, in the long run, proved fatal to the idea that logical positivism described the actual structure of theory development: It simply does not work that way. For
instance, in his historical overview of social science measurement, Wright (1997) quotes a passage from the Magna Carta, dating from 1215, in which King John of England declared:

“There shall be one measure of wine throughout our kingdom, and one of ale, and one measure of corn, to wit, the London quarter, and one breadth of cloth, to wit, two ells within the selvages. As with measures so shall it be with weights.”

The quotation illustrates the obvious but interesting fact that King John, who clearly understood the basic principles of measurement and their importance quite well, is nevertheless unlikely to have thought of measurement as the homomorphic mapping of an empirical relational system into a numerical one. Likewise, carpenters, tradespeople, and scientists use the principles of measurement without apparent awareness of the higher mathematics they are involved in. So, unless we assume that Freud's unconscious not only exists, but actually has a serious expertise in set theory, it would surely be outrageous to say that representational measurement theory is a theory of how measurement is carried out in practice. King John knew nothing about set theory, and still he believed that measurement was important enough to include a passage about it in the Magna Carta. The Egyptians certainly knew how to measure the bricks they used in constructing pyramids, but they did not even have the number zero, let alone the real number system. Scientists use measurement procedures all the time, but they cannot explain to you how representational measurement theory works unless they have studied it. Clearly, the reconstruction given is not a reconstruction of the historical development of measurement, and neither can it be interpreted as a psychological description of the scientist carrying out the measurement procedure.

The question that forces itself upon us then becomes: If representationalism offers a reconstruction, then what is it reconstructing? Exactly the same problem was faced by the logical positivists, who described the structure of scientific theories in a way that was quickly realized to be inadequate as an actual description. Reichenbach (1938) circumvented this problem by stating that he was giving a rational reconstruction of scientific theories. Such a conceptualization seems to fit representationalism quite nicely. In accordance with the findings in the previous section, we may then interpret the theory not as showing how a scale can be constructed, but how a scale could be constructed by a Laplacean demon, i.e., a being equipped with powers that enable him to make errorless observations of qualitative relations between objects. I think this is the best possible interpretation of the theory. Representationalism offers a theory of homomorphisms that certainly has intuitive appeal in its mathematical elegance and parsimony. The interpretation in terms of rational reconstruction does nothing to devaluate the theory at this level. However, the interpretation also makes clear that representationalism is not a theory about how the concept of measurement developed, and it is not a theory of what scientists do, for the simple reason that they are not Laplacean demons. While this does not diminish the importance of the analysis for understanding measurement structures, it does raise doubts with respect to the prescriptive force of representationalism.
Reconstruction does not entail prescription

One of the most serious mistakes one can make when thinking about any research topic is the following: in a certain population, we have observed that elements with property $x$ tend to have property $y$, so if we equip elements lacking property $x$ with that property, then they will also develop property $y$. Philosophers of science who view it as their mission to equip scientists with prescriptive criteria of theory development, model selection, or scale construction, commonly commit this fallacy. For example, they think that, because successful scientific theories have property $x$, we can construct successful theories by forming theories that have property $x$.

So, because we observe that Copernican astronomy is more parsimonious and yet equally adequate in prediction as the Ptolemean system, we should select models in psychology on the basis of parsimony and predictive adequacy—regardless of whether such a strategy has been shown to yield better theories in psychology (it has not). Because we observe that highly successful theories in physics, like Newtonian mechanics and the theory of relativity, are falsifiable, we should construct falsifiable theories in psychology, and psychology will automatically become a serious scientific discipline—regardless of whether it makes any sense to construct falsifiable theories at the present stage of theory formation in psychology. And so it is with the advocates of fundamental measurement: Measurement has been very successful in physics, where it allegedly obeys the structure of fundamental measurement theory, so if we construct psychological measures based on this theory, then psychology will finally become the long-sought quantitative science we have all been dreaming of. This kind of science fiction is continually being propagated by the advocates of Rasch models (Wright, 1997; Bond & Fox, 2001), fed by the otherwise admirable theoretical work of Michell (1990, 1999, 2000), and has led Kline (1998) to adopt the mysterious view that psychology cannot be scientific without fundamental measurement.

This puts the horse behind the cart. We have seen that fundamental measurement theory cannot be interpreted as more than a rational reconstruction of the measurement process. This does not devaluate the theoretical insights it provides, but it is important to keep in mind that reconstruction does not entail prescription. One should remember that fundamental measurement theory is entirely post-hoc. It has not helped to construct measurements of length or mass; it has not even helped to construct measurements of conjoint concepts like force or density. It has elucidated the structure of such cases of measurement. But it is highly questionable, given its deterministic nature, whether the theory could have been of any use in constructing these measures in the first place. Probably, the clean logic of the theory would have been an obstacle rather than an aid to the development of measurement procedures—a process that is messy, inexact, full of spurious results, packed with error variance, and that more often than not requires decisions to be made on the basis of intuition than on the basis of logic. It is a fallacy to think that, because established forms of measurement allow for a philosophical reconstruction in terms of model $x$, all measurements should be constructed to obey the prescriptions of model $x$—regardless of whether model $x$ is a fundamental measurement model, a latent variable model, a generalizability theory model, or some other technique.
The reason why fundamental measurement theory cannot be a prescriptive methodological framework has to do with its particular deterministic structure, but much more with the fact that it is just a model. One should be weary of models that are propagated as prescriptive frameworks in a universal sense, because whether or not a model is adequate in a given situation strongly depends on the substantive considerations that are relevant in that particular situation. Additivity, for example, is desirable because it is simple, but it is only desirable if substantive theory suggests that additivity should hold. Now, if we had different kinds of meter sticks that produced crossing interactions, we would certainly be surprised: It would be strange if on some meter sticks longer things had a higher expected value than shorter things, and on other meter sticks longer things had a lower expected value than shorter things. Yet, this is what the presence of crossing interactions (as present in factor models with unequal factor loadings as well as in Birnbaum models) signifies. On the other hand, if we asked people a number of questions to measure their height, we might certainly encounter such situations. For instance, the item “I can touch the top of a doorpost with my hands” can reasonably be considered to measure bodily height, be it indirectly. It will show a quite steep curve as a function of height, jumping from “no” to “yes” at about 1.75 meters. Coding the item as “yes”:1 and “no”:0, we might imagine this item to have an expected value of .80 for people 1.80 meters tall, and an expected value of .20 for people 1.70 meters tall. The item “I am pretty tall” is less direct, but may nevertheless be considered to validly measure the trait at hand. Because it is less direct, the item characteristic curve will not jump from 0 to 1 as suddenly and steeply as the previous item. This yields the possibility that people who are 1.70 meters tall will have an expected value of .30, while people who are 1.80 meters tall may have an expected value of .70. Thus, for people who are 1.80 meters tall, the first item is ‘easier’ than the second, but for people who are 1.70 meters tall, the second item is easier than the first. Technically, this means that there is a crossing interaction between the subject and item factors, which implies that additivity is violated and no conjoint additive representation can be found. Does this mean we cannot use the two items to construct a valid measure for height? And what about items used to measure cognitive abilities or personality characteristics? Should we always demand additivity in such cases? One should be weary to draw this conclusion because it depends on a dogmatic view that leans towards essentialism about the term ‘measurement’. In the absence of a rationale based on substantive, rather than philosophical, considerations that sustain various formal properties like additivity (or, for that matter, unidimensionality, measurement invariance, and the like) one should be very careful in propagating the universal demand for such properties. It amounts to pure speculation to say that constructing measures on the basis of these formal criteria will lead to better measurement.

It is certainly true that many measurement practices in psychology, as well as models that employ continuous latent variables, assume that psychological attributes are quantitative. This is a serious assumption, as has been clarified by Michell (1990; 1999). It is also true that finding an additive conjoint measurement representation, even if it is probabilistic like in the Rasch model or the ADISOP models discussed by Scheiblechner (1999), yields support for this assumption. But
it is not true that the existence of a continuous, quantitative psychological attribute guarantees that one can construct a psychological test for it that allows for an additive representation. It has never been proven, and will never be proven, that constructing tests on the basis of this requirement will yield better tests, regardless of the dimension one is trying to measure or the substantive field one is working in.

In the substantive context of psychology, we should seriously consider the possibility that psychological measurement is so much more complicated than measurement in physics, that it is a different ball game altogether. That is, fundamental measurement may be fundamentally inappropriate. Consider the simple fact that some questions are better asked in a dichotomous format, while some perform better in a polytomous format, whereas still other questions are more appropriately put in open-ended form. Assume for a moment that the trait indicated by the term ‘general intelligence’ exists, and is quantitative. It may well be that the best possible test for intelligence would be composed of different item formats (in fact, such tests commonly are). Now how are we going to implement the constraints of fundamental measurement in this case? Clearly, we are not going to do this at all, at least not in the presently available forms of the model.

Now one may either give up and conclude that psychology “will never be a science” (Kline, 1998), or one may try to accommodate for the problem, for example by assuming that a latent variable exists and underlies the observations, and try to build a model that can handle the situation. A serious attempt at doing this has been undertaken by Moustaki & Knott (2000), who have formulated a general latent variable model that allows for the use of different item formats. Probably, this model would be rejected by fundamental measurement theorists: No representation or uniqueness theorems are available for it, and I would not be surprised if it were proven that no appropriate representations exist at all. Therefore, the model cannot be said to be a measurement model in the fundamental measurement theory sense. But is this the constructive contribution of representationalism to the problem of psychological measurement? That it is impossible to measure something if it does not allow for a formal reconstruction in terms of representation and uniqueness theorems? I consider that a very unsatisfying option. It is undoubtedly the case that representationalism has done much to help the case of psychological measurement, and there may indeed be applications of the theory, in psychophysics and other basic areas of psychology, where it actually works. In the case of higher order constructs like extraversion, intelligence, and attitudes, however, the theory has not been very useful so far. And considerations like the above (and there are many more) suggest that it will not do very much in these areas. The idea that the representationalist strategy is required in every case of measurement may be taken to be a truism, if one limits the meaning of measurement to homomorphic representation. I consider such a limitation to be unduly restrictive. Moreover, because representationalism is no more than a reconstruction, the general demand for fundamental measurement is based on an overinterpretation of the theory.

It seems that the advocates of a prescriptive reading of fundamental measurement theory are making a serious mistake. They think that, because physical measurement can be reconstructed in terms of axiomatic theory, it follows that psychology should construct psychological tests through the application of axiomatic
4.2 Three perspectives on measurement scales

theory. The justification for this conviction is usually cast in terms of logical, philosophical, and mathematical arguments, but as it stands it is no more than a belief. Moreover, the arguments adduced to support it are based on the fallacy that something that works in one scientific field (physics) will also work in a completely different field (psychology). Arguments for or against a psychological method, however, should be based on considerations that bear on psychology, and not on considerations that bear on physics. I congratulate the physicists with the lucky situation that the structure of the world they study is simple enough for them to reconstruct measurement procedures in terms of the lucid and powerful formal framework of fundamental measurement. However, in psychology we do not study stones, atoms, and quarks, but human beings. The human being is one of the most complex systems in the universe. Still, with respect to some psychological characteristics, human beings seem to vary from one another in a systematic fashion, and psychological measurement procedures attempt to capture this variation. Everybody knows that these practices are packed with assumptions, both substantial and auxiliary, and that the measurement outcomes they return should be interpreted with care. These measurement outcomes will more often than not violate the requirements of fundamental measurement. But to say that this implies that they are not measurements at all, and that the only way to construct measurement is to follow the axiomatic theory, is preliminary and hinges on an essentialist philosophy concerning the meaning of the term ‘measurement’.

It is clear that, if one chooses to define measurement as homomorphic representation, then many assessment techniques that are commonly viewed as instances of measurement are not covered by this definition. So interpreted, these procedures do not yield measurements. If it makes the fundamental measurement theorists feel better, we may decide to call them ‘fmeasurements’ instead – although I doubt whether that will make much of a difference. However, it is not a fact, either empirical, logical, mathematical, or otherwise, that measurement is homomorphic representation. At most, it is a convention. If we decide to designate by the term ‘cow’ every animal with eight eyes, then what we commonly recognize as being a cow no longer is, while spiders are suddenly in the possession of cowhood. While this will change the meaning of the word ‘cattle’ as well as the size of T-bone steaks, this juggling around of terms will not add much to the science of biology. The same holds for measurement in psychology. Nothing forces a specific definition of the term upon us, and nothing forces us to follow a specific approach towards psychological testing; certainly not when we consider the observation by Cliff (1992) that the axiomatic approach has not been able to produce a single striking psychological example to illustrate its benefits. Thus, it would perhaps be an idea for the advocates of the prescriptive reading of fundamental measurement theory to start showing the superiority of the approach, rather than to talk about it. For the prescriptive reading of the theory is not founded upon a serious consideration of the problems inherent in psychological measurement, but rather on a mindless mimicking of physics. In this context, it is ironic that the theory that originated as a reply to the overstated conclusions of Campbell and his associates has, in the hands of some of the more vigorous proponents of fundamental measurement, come to occupy the very same position it once sought to oppose.
4.3 Discussion

Representationalism offers a powerful conceptual framework for thinking about measurement. The focus on mapping empirical relations into a number system has liberated measurement theory from the demand that a concatenation operation must always be available, and as such it has provided a justification for moving away from operationalism. The logic of the theory allows one to see very clearly what one is assuming in different situations, and this is indeed an invaluable theoretical aid in model construction and evaluation. One should, however, not overinterpret the theory. As a formal framework, the theory can be considered adequate for its purpose, but as a philosophical or even prescriptive framework, it is too simplistic. In a philosophical interpretation, the clean logic that is the theoretical strength of the model quickly becomes its weakness. Representational theory has no natural means of incorporating error, and must abandon its central tenets when it is equipped with a method to do so. In face of this problem, a choice has to be made between two conclusions: Either measurement is impossible in the presence of error, or representational measurement theory is not a theory of how measurement is, can be, or should be carried out in practice. The first conclusion is absurd, but rejecting it leads immediately to the second, and this raises the question what the status of representationalism is. It has been argued here that representationalism offers a rational reconstruction of the measurement process. That is, it elucidates measurement procedures by recasting them in idealized logical terms, and it does this very well.

Whatever the conceptual status of rational reconstruction may be, however, it does not have prescriptive force. Therefore, the advocates of a prescriptive reading of the theory are not justified in their position. In effect, they are trying to sell a conceptual theory of measurement as a method for test construction and analysis. A method, however, can only be used if applicable, and because the inability to deal with imperfect observations is so deeply entrenched in the structure of representational theory, its applicability in the social sciences must be considered limited. It seems safe to assert that, in psychology, the clean observations that representationalism requires will not be realized in our time, if at all. Therefore, it is unclear what theorists like Michell (1990; 1999; 2000) and Kline (1998) are advocating. At best, they must be interpreted as proposing the use of unidimensional, additive IRT models, because these are the only models that allow for error and bear at least a superficial resemblance to additive conjoint structures. So interpreted, however, their claims do not seem all that radical. Moreover, the step from additive to non-additive latent variable models, while philosophically important, is a small one from statistical, practical, and substantive points of view. Substantive considerations do not generally support additivity, as is evidenced by the fact that it is easy to give examples of test items that violate it but still measure the same latent variable. Statistical considerations suggest that the odds of finding a model with perfectly parallel item response functions are vanishingly small, so that the demand should not be taken overly seriously. Practical considerations lead to the conclusion that it will be virtually impossible to construct a situation, where a model that satisfies the usual IRT assumptions (monotonically increasing item response functions, uni-
dimensionality, and local independence) fits the data, but the correlation between the simple sumscore and the latent variable drops below .90, which would seem more than enough for the average researcher. We must therefore conclude that positing additivity as a universal demand is, at best, preliminary.

Moreover, from a philosophical viewpoint, the difference between additive and nonadditive latent variable models seems much smaller than the difference between latent variable models and strict representations. A researcher working within representationalism is not making claims with respect to the question where the data came from - he is merely representing the data. From this viewpoint, additivity is central, because a violation of additivity precludes the possibility of homomorphic representation using a quantitative metric. The researcher who uses latent variable theory is engaged in a different activity. He knows (or should know) that homomorphic representation is strictly taken impossible because he is modeling stochastic relations that are not directly observable. In order to model such relations, he posits the existence of a latent variable. Because he is now introducing metaphysics, he needs a justification for these metaphysics, and this justification will not come from the data or from methodology. For instance, what type of latent variable model to use (e.g., a class model or a trait model), or how to conceptualize the relation between the latent variable and the observed variables (e.g., as additive or nonadditive, parametric or nonparametric), are typical examples of questions that cannot be answered by the data or by methodological considerations. Exactly because the researcher is now positing a data generating mechanism, rather than constructing a representation, these questions must be answered by substantive theory. For instance, developmental theory suggests that conservation data (Dolan, Jansen, & Van der Maas, *submitted*) should not be modeled as originating from a continuous latent variable model, but from a class model where the classes correspond to different developmental stages. Nothing in the data themselves forces this choice; it may well be possible to model the data using a continuous latent space. Neither do there exist methodological considerations that say developmental data should be modeled in this particular way. Clearly, the burden of proof shifts from the area of logic and mathematics to the area of substantive theory, which must give a justification for the metaphysics introduced.

Thus, a researcher, who conceptualizes a psychological construct as a measurement scale, is ascribing a completely different theoretical status to that construct as compared to a researcher who conceptualizes a construct as a latent variable. A measurement scale is a representation of observed relations, whereas a latent variable model is a guess about the structure of the data generating mechanism, i.e., a posited probabilistic explanation of such relations. Mathematically, the representationalist approach is as useful for studying latent variable models as for studying deterministic measurement structures. Its focus on mappings yields interesting insights into the relations between various levels of representation, and is a good theoretical tool in the study of different models and the relations between them. This holds true for the latent variable model as it does elsewhere. These insights, however, have more to do with the formal logic of the theory than with its philosophical account of what measurement is. The philosophical account of representationalism involves a highly restrictive empiricist point of view. It requires the
direct observation of relations among objects in every case - the conjoint structures included. Representationalism could be considered to play an important role in empiricism, because it gives an account of how we get from qualitative observations to quantitative theoretical terms. This avoids metaphysical speculation, because the observations can be conceptualized as being made with the unaided eye. However, if we could judge the relation 'more intelligent than' directly, there would be no need for intelligence tests. From this point of view, the problem in psychological measurement is simply that the unaided eye does not work very well. It has to be supplemented by statistical assumptions concerning the behavior of aggregates, substantive hypotheses on the nature of data generating processes, and metaphysical postulates concerning the existence of propensities and latent variables. All this is required in order to get the endeavor off the ground in the first place. Representationism does not pay attention to these problems but ignores them. This is fine as long as the concerns of the theory are limited to formal structures, but when interpreted as a conceptual framework for tackling the problem of psychological measurement in general, or even as a prescriptive framework for scale development, the theory can hardly be considered adequate. Thus, although representationism is very important in elucidating some of the problems in psychological measurement, the mathematical structure of models, and the differences between measurement in the natural sciences and in psychology, its importance is limited and should not be overstressed.